SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 1, 1960

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Tokyo

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 1, 1960

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Tokyo

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

LIST OF CONTENTS

NUMBER 1

Foreword 1

L. A. D'Asaro: Diffusion and oxide masking in silicon by the box method . . 3

L. Esaki and Y. Miyahara: A new device using the tunneling process in narrow *p–n* junctions 13

R. A. Pucel: Physical principles of the Esaki diode and some of its properties as a circuit element 22

U. Birkholz: Zur experimentellen Bestimmung der thermoelektrischen Effektivität von Halbleitern 34

T. Renner: Die thermodynamischen Daten und Dampfdrucke der Wichtigsten III–V-Verbindungen 39

J. Oroshnik and A. Many: Quantitative photovoltaic evaluation of the resistivity homogeneity of germanium single crystals 46

A. F. Gibson and J. R. Morgan: Avalanche injection diodes 54

C. A. Hogarth: A method of making negative-resistance two-state switching diodes . 70

G. Arlt: Halleffekt-vierpole mit hohem Wirkungsgrad 75

M. Green and I. N. Greenberg: Transient charge-carrier diffusion in high-purity silicon 85

Note:
F. E. Roberts: Preferential plating on germanium 93

Book Review:
Masers, by G. Troup (Reviewed by P. N. Butcher) 94

Notice 95

Papers to be published in future issues 95

NUMBER 2

Editor's Foreword i

F. A. Cunnell, J. T. Edmond and W. R. Harding: Technology of gallium arsenide . 97

J. M. Lavine: The behavior of *p*- and *n*-doped contacts in a space-charge depletion region 107

R. Saintesprit: Sur la diffusion des éléments des colonnes III et V dans le silicium . 123

A. J. Strauss, T. C. Harman, E. B. Owens and M. C. Finn: Indium as a source of impurities in indium arsenide 131

J. E. Parrott: The interpretation of the stationary and transient behaviour of refrigerating thermocouples 135

R. L. Williams: Radiation limited PbS cells 144

R. L. Williams: Minority-carrier effects in chemically deposited PbS photoconductive films 148

Note:
W. Rindner: A masking technique for vacuum evaporation 157

NUMBER 3

MOTOSUKE MUNAKATA: Electrical conductivity of high vanadium phosphate glass . . 159

M. F. LAMORTE: Calculation of concentration profiles and surface concentration from sheet-conductance measurements of diffused layers 164

JULIUS BRUNNER: Der Halleffekt im inhomogenen Magnetfeld 172

S. DUINKER and B. VAN OMMEN: The scansor, a new multi-aperture rectangular-loop ferrite device 176

I. A. LESK and H. A. JENSEN: Germanium *p–n*-junction–tunnel-junction combination devices 183

C. H. L. GOODMAN: The regrown-diffused transistor 188

M. A. LEE: Variation with temperature of the distribution coefficient of indium in germanium 194

J. A. M. DIKHOFF: Cross-sectional resistivity variations in germanium single crystals . 202

C. A. MEAD: The operation of junction transistors at high currents and in saturation . 211

H. WEISS: Zur Messung von magnetischen Feldern mit Hallgeneratoren 225

N. J. HARRICK: Semiconductor type and local doping determined through the use of infrared radiation 234

Notes:

J. C. BRICE and A. A. STRIDE: A continuous-reading four-point resistivity probe . 245

A. C. ENGLISH and R. W. THOMAS: Creep and forming in selenium rectifiers . . 245

Book Reviews:

Principles of Semiconductor Device Operation, by A. K. JONSCHER (Reviewed by ALAN F. GIBSON) 250

Electrical Noise. Fundamentals and Physical Mechanism, by D. A. BELL (Reviewed by L. PINCHERLE) 250

Erratum:

M. GREEN and I. N. GREENBERG: Transient charge-carrier diffusion in high-purity silicon 252

NUMBER 4

SUPERCONDUCTIVE TECHNIQUES

Editors' Foreword 253

B. SERIN: Outline of recent developments in superconductivity 255

I. M. TEMPLETON: The use of superconductive devices in research at low temperatures 258

V. L. NEWHOUSE, J. W. BREMER and H. H. EDWARDS: Characterization and physics of the crossed-film cryotron—A Review 261

H. O. MCMAHON and W. E. GIFFORD: Closed-cycle helium refrigeration 273

C. R. VAIL, M. S. P. LUCAS, H. A. OWEN and W. C. STEWART: An approach to the experimental study of persistent-current devices 279

W. H. CHERRY and J. I. GITTLEMAN: Thermal and electrodynamic aspects of the superconductive transition process 287

D. H. PARKINSON: British research on the Crowe cell 306

NUMBER 4—*continued*

P. R. STUART: British research on thin-film cryotrons 312

R. F. BROOM and E. H. RHODERICK: A new type of bistable element involving thermal propagation of a normal region in a thin superconducting film 314

J. D. BLADES, J. GERBER and C. T. THOMPSON: Initial studies of the preparation and properties of vanadium thin films 319

F. W. SCHMIDLIN, A. J. LEARN, E. C. CRITTENDEN, JR. and J. N. COOPER: Current-induced switching of superconductive thin films 323

R. E. GLOVER, III: High-frequency behavior of very thin superconducting films and Changes in critical temperature produced by electrostatic charging 335

R. E. JACOBSSON: Research on superconductive devices in Sweden 339

D. ABRAHAM: Superconducting films less than 100 Å thick 340

L. L. BURNS, G. W. LECK, G. A. ALPHONSE and R. W. KATZ: Continuous-sheet superconductive memory 343

M. L. COHEN and J. L. MILES: Characteristics of thin-film cryotrons 351

A. E. SLADE and C. R. SMALLMAN: Thin-film cryotron memory 357

A. M. TOXEN: Characteristics of bulk and thin-film superconducting alloys . . 363

D. P. SERAPHIM: Effect of defects on the superconducting properties of tantalum . . 368

D. R. YOUNG, J. C. SWIHART, S. TANSAL and N. H. MEYERS: Use of a superconducting transmission line for measuring penetration depths 378

R. B. DELANO, JR.: Edge effects in superconducting films 381

G. B. ROSENBERGER: Operation and analysis of planar cryotrons and simple cryotron circuits 388

M. K. HAYNES: Cryotron storage, arithmetic and logical circuits 399

AUTHOR INDEX

(B.R. indicates a Book Review)

ABRAHAM D., 340
ALPHONSE G. A., 343
ARLT G., 75
BIRKHOLZ U., 34
BLADES J. D., 319
BREMER J. W., 261
BRICE J. C., 245
BROOM R. F., 314
BRUNNER J., 172
BURNS L. L., 343
BUTCHER P. N., 94 (B.R.)
CHERRY W. H., 287
COHEN M. L., 351
COOPER J. N., 323
CRITTENDEN E. C., JR., 323
CUNNELL F. A., 97
D'ASARO L. A., 3
DELANO R. B., JR., 381
DIKHOFF J. A. M., 202
DUINKER S., 176
EDMOND J. T., 97
EDWARDS H. H., 261
ENGLISH A. C., 245
ESAKI L., 13
FINN M. C., 131
GERBER J., 319
GIBSON A. F., 54, 250 (B.R.)
GIFFORD W. E., 273
GITTLEMAN J. I., 287
GLOVER R. E., III, 335
GOODMAN C. H. L., 188
GREEN M., 85, 252
GREENBERG I. N., 85, 252
HARDING W. R., 97
HARMAN T. C., 131
HARRICK N. J., 234
HAYNES M. K., 399
HOGARTH C. A., 70
JACOBSSON R. E., 339
JENSEN H. A., 183
KATZ R. W., 343
LAMORTE M. F., 164
LAVINE J. M., 107
LEARN A. J., 323
LECK G. W., 343
LEE M. A., 194
LESK I. A., 183
LUCAS M. S. P., 279
MCMAHON H. O., 273
MANY A., 46
MEAD C. A., 211
MEYERS N. H., 378
MILES J. L., 351
MIYAHARA Y., 13
MORGAN J. R., 54
MUNAKATA M., 159
NEWHOUSE V. L., 261
OMMEN B. VAN, 176
OROSHNIK J., 46
OWEN H. A., 279
OWENS E. B., 131
PARKINSON D. H., 306
PARROTT J. E., 135
PINCHERLE L., 250 (B.R.)
PUCEL R. A., 22
RENNER T., 39

AUTHOR INDEX—*continued*

Rhoderick E. H., 314
Rindner W., 157
Roberts F. E., 93
Rosenberger G. B., 388
Saintesprit R., 123
Schmidlin F. W., 323
Seraphim D. P., 368
Serin B., 255
Slade A. E., 357
Smallman C. R., 357
Stewart W. C., 279
Strauss A. J., 131
Stride A. A., 245
Stuart P. R., 312
Swihart J. C., 378
Tansal S., 378
Templeton I. M., 258
Thomas R. W., 245
Thompson C. T., 319
Toxen A. M., 363
Vail C. R., 279
Weiss H., 225
Williams R. L., 144, 148
Young D. R., 378

FOREWORD

WITH the invention of the transistor and the discovery of the microwave applications of ferrites about ten years ago, a tremendous upsurge of interest has occurred in the applications of solid-state physics to electronic technology. Recent years have seen a great proliferation of solid-state devices which have revolutionized the field of electronics and have led to the growth of an entirely new grouping of scientists and engineers throughout the world. It seems likely that what we see now is only a beginning. Thus the launching of a new international journal, *Solid-State Electronics*, to serve this new world-wide audience seems especially opportune at this time.

Stressing as it does solid-state device theory and technology, the new journal provides an excellent complement to the three-year-old *Physics and Chemistry of Solids*, which is concerned with the more fundamental and theoretical aspects of solid-state science. As Editor of the latter journal it gives me real pleasure to welcome this newcomer, and to take this occasion to express my hearty endorsement of its scope and aims.

HARVEY BROOKS

BLANK
PAGE

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 3–12. Printed in Great Britain

DIFFUSION AND OXIDE MASKING IN SILICON BY THE BOX METHOD

L. A. D'ASARO

Bell Telephone Laboratories, Murray Hill, New Jersey

(*Received* 20 *May* 1959)

Abstract—The "box method" for solid-state diffusion into silicon uses an impurity oxide dissolved in SiO_2 as a source of diffusant within a loosely closed box. The surface concentration of the impurity in the silicon can be varied over several orders of magnitude by a variation of the source composition. The method is well adapted for the use of a silicon oxide layer as a mask against diffusion. The masking action is particularly effective if the impurity oxide concentration in the source is below the value which produces a liquid.

Specific experimental results are reported for the diffusion of boron and phosphorus. For example, oxide masking against diffusion was obtained in 0·3 Ω-cm silicon under a condition which gave in unoxidized regions a diffused layer with a surface concentration greater than 10^{20} atoms/cm^3 and a junction depth of 3×10^{-4} in.

Résumé—La méthode de la "boite" pour effectuer des diffusions dans le silicium, utilise comme source de diffusant une solution d'un oxyde de l'impureté dans SiO_2, à l'intérieur d'une boite presque fermée. On peut varier la concentration des impuretés dans le silicium sur plusieurs ordres de grandeur, en variant la composition de la source. La méthode est bien adaptée à l'utilisation d'une couche d'oxyde de silicium comme "masque" contre la diffusion. L'effet "masque" est particulièrement efficace si la concentration de l'oxyde d'impureté dans la source est inférieure à la valeur pour laquelle un liquide apparait.

Des résultats expérimentaux sont présentés pour le cas de la diffusion du bore et du phosphore. Par exemple, on peut obtenir un effet de "masque" contre la diffusion dans du silicium de 0,3 Ω-cm dans des conditions où dans les régions non-oxydées on obtient une couche diffusée de concentration de surface supérieure à 10^{19} atomes/cm^3 et une profondeur de jonction de 3×10^{-4} in.

Zusammenfassung—Die „Gefässmethode" für Festkörperdiffusion in Silizium verwendet ein in SiO_2 aufgelöstes Störstellenoxyd als Diffusionsquelle innerhalb eines geschlossenen Gefässes. Die Oberflächenkonzentration der Störstellen im Silizium kann durch eine Variation der Quellenzusammensetzung um einige Grössenordnungen abgewandelt werden. Die Methode ist für die Verwendung einer Siliziumoxydschicht als Maske gegen Diffusion gut geeignet. Die maskierende Wirkung ist besonders wirkungsvoll, wenn die in der Quelle befindliche Konzentration des Störstellenoxyds unter dem Wert liegt, der eine Flüssigkeitsstruktur erzeugt.

Versuchsergebnisse werden für die Diffusion von Bor und Phosphor berichtet. So wurde z.B. oxydische Maskierung gegen Diffusion in 0,3 Ω-cm Silizium erhalten unter einer Bedingung, die in nichtoxydierten Bereichen eine diffundierte Schicht mit einer Oberflächenkonzentration grösser als 10^{20} Atome/cm^3 und einer Tiefe des p–n-Überganges von 3×10^{-4} in. ergab.

1. INTRODUCTION AND PRINCIPLE

THE formation of p–n junctions in silicon by solid-state diffusion of impurities can be carried out by a variety of methods.[1] Diffusions performed at atmospheric pressure are especially convenient. Most published reports on such techniques follow the method initially described by FROSCH and DERICK[2]. They showed that at atmospheric pressure the use of an oxidizing gas prevents erosion of the silicon, and that a layer of SiO_2 on the silicon can act as a mask against diffusion of impurities. In their arrangement an impurity vapor was introduced by passing the oxidizing gas over the heated impurity and then over the sample.

In the present method the diffusion is carried out in a closed box (Fig. 1) in which impurity is introduced from a solution of impurity oxide in SiO_2. The vapor pressure of the impurity under near-equilibrium conditions is controlled by dilution of

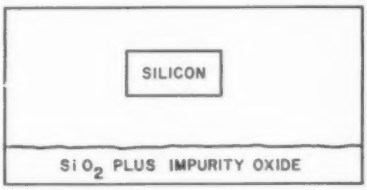


FIG. 1. Diagram of a closed box containing a solution of impurity oxide in SiO_2 and a sample of silicon to be diffused.

the impurity source. At the surface of the SiO_2 layer on the silicon, the vapor pressure of impurity will result in an impurity oxide concentration which is close to the impurity oxide concentration in the source. If no SiO_2 layer is originally present on the silicon, the impurity will be incorporated into a newly formed SiO_2 layer. The impurity oxide will react with the silicon to give rise to a diffused distribution of impurity. On the other hand, if a previously grown SiO_2 layer is present, the impurity must penetrate the SiO_2 layer before diffusion into the silicon is possible. With a thick SiO_2 layer, the penetration can be sufficiently slow so that masking is obtained.

The most effective masking results when the rate at which the impurity penetrates the SiO_2 layer is minimized. Since one expects that the rate of penetration will decrease with increasing viscosity of the SiO_2 layer, and since the state of the source material is a guide to the state of the surface of the SiO_2 layer, it follows that the best masking is expected using a source which is solid at the diffusion temperature. In order to obtain a solid source, the impurity concentration must be sufficiently small, since a decrease of impurities results in an increase of the viscosity. It will be shown in the following that the consequent decrease in surface concentration can be compensated to a certain extent by a decrease in oxygen partial pressure.

The following section is concerned with a more detailed analysis of such a system. First an idealized arrangement will be discussed, in order to emphasize the general features. Then the deviations from ideality in a practical system will be analyzed.

A final section contains experimental arrangements and results.

2. THEORY

(A) *The interaction of source and sink in an idealized system*

An ideal arrangement for the diffusion would be a perfectly tight box containing a controlled atmosphere and a partial pressure of impurity oxide, both of which remain constant. The impurity oxide in the vapor phase provides a means of interaction between the source material and the silicon sample (the sink).

The source material at the diffusion temperature establishes an equilibrium vapor pressure of impurity oxide. This vapor pressure is determined by the concentration of impurity oxide in the source. If the solution were ideal in the chemical sense, there would be a linear reduction in vapor pressure with dilution,

$$P_A = F_A P_A^0 \tag{1}$$

where P_A is the vapor pressure of component A, P_A^0 is the vapor pressure of the pure substance, and F_A is the mole fraction of component A. In a nonideal solution there is an interaction between the atoms of solvent and solute which makes the vapor pressure depart from the value given by equation (1). In any case, the equilibrium vapor pressure is a unique function of the dilution.

Using published values of vapor pressure of B_2O_3, for example, and information on the penetration of impurities through the oxide as obtained in this work, one can estimate that the times to reach equilibrium vapor pressure are of the order of seconds. These times are always small fractions of the total time for a diffusion run.

When a small piece of oxidized silicon is placed inside the box, the SiO_2 layer on the silicon acts as a sink for the gaseous impurity oxide. Under equilibrium the surface of the SiO_2 layer will achieve the same composition as the source material. The impurity oxide at the surface of the SiO_2 layer will act as a source for diffusion into the SiO_2. Somewhere between the SiO_2–gas interface and the SiO_2–Si interface the impurity oxide will be reduced. Part of the elemental impurity will

diffuse into the silicon. If R represents a chemical impurity and RO its oxide, the reduction process will proceed by a reaction such as

$$2RO + Si \rightleftarrows SiO_2 + 2R. \quad (2)$$

The possibility that the source of elemental impurity is the dissociation reaction

$$2RO \rightleftarrows 2R + O_2 \quad (3)$$

may be ruled out by a consideration of the free-energy change ΔF for reaction (2) as compared with reaction (3).[3] In the case of B_2O_3, for example, ΔF for reaction (2) is 3300 cal/mole, while for reaction (3), $\Delta F = 430{,}000$ cal/mole. These values lead to an equilibrium constant of 0·3 for reaction (2), but only 2×10^{-60} for reaction (3).

While reaction (2) is proceeding the silicon is oxidizing,

$$Si + O_2 \rightleftarrows SiO_2. \quad (4)$$

If the oxygen concentration is decreased, reaction (4) within the oxide layer is shifted towards the left. The increase in concentration of silicon then operates to shift reaction (2) towards the right, thus increasing the concentration of the elemental impurity R.

As mentioned before, the effectiveness of an SiO_2 layer used as a mask will be greatest when the amount of impurity oxide dissolved in it is least. For larger impurity oxide concentrations there is a decrease of viscosity of the glassy SiO_2–impurity oxide solution. Although such solutions have no well-defined melting point, a phase diagram such as shown in Fig. 2* provides a rough guide to the state of the solution. Since the state of the source material can be taken as a guide to the state of the surface of the SiO_2 layer, the state of the source is a measure of the expected effectiveness of masking.

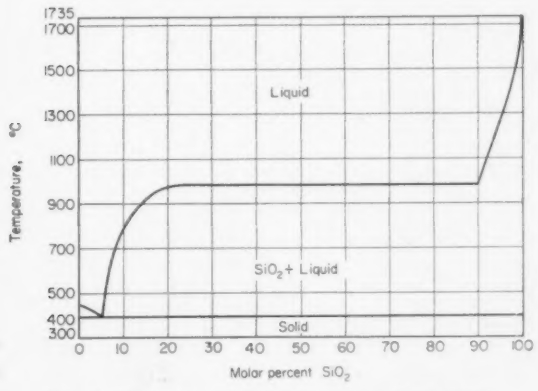


FIG. 2. Phase diagram for the system B_2O_3–SiO_2.

From the previous discussion it follows that a decrease of impurity oxide concentration in the source causes a decrease in surface concentration of the diffused layer. This decrease can be compensated by a decrease in the oxygen partial pressure, which causes an increase in surface concentration. The experimental results to be given later show that with boron and phosphorus the solid solubility of these elements can be approached using sources which are still quite viscous. Hence effective masking is possible without a decrease in surface concentration.

(B) *Effects expected in a practical system*

Although a completely sealed box would be the best arrangement for approaching equilibrium conditions, it is not convenient to seal completely. A box with a finite leak rate can be designed to be opened or closed easily. Furthermore, such a leak both avoids complications due to expansion of the gas at high temperatures and facilitates the communcation of a surrounding atmosphere into the interior of the box. Fig. 3 shows a diagram of a small box or capsule made of quartz and closed with a platinum band. This design was used for the experiments to be described later.

The leak to the outside of the capsule in Fig. 3 is a long narrow duct. For the purposes of analysis such a capsule can be represented by the idealized box of Fig. 4, with an opening through a duct of cross-sectional area A_d and height h. The leak rate through the duct results in a loss of impurity vapor and a gradual depletion of the source. The associated vapor pressure therefore decreases. It is desired that this fall-off in vapor pressure should be minimized during a diffusion run.

In order to analyze the rate of depletion of the source, three factors should be considered:

(i) The diffusion of impurities out of the source.
(ii) A possible rate process at the surface of the source.
(iii) The pumping speed of the duct.

* The phase diagram was derived from a three-phase diagram given by HALL and INSLEY[4].

The effect of these factors is considered in the Appendix where a numerical estimate of the rate of depletion is given for a practical case. It is

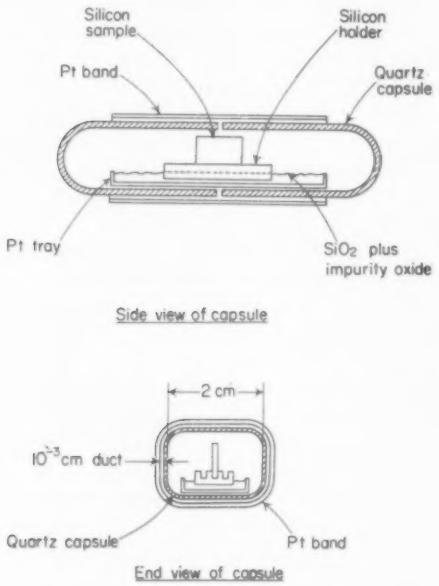


FIG. 3. Quartz capsule closed with a platinum band and containing source material and sample to be diffused.

shown there that the rate of depletion can be kept so low that 70 hr of heating at 1300°C would be required to deplete a low-concentration B_2O_3 source by 10 per cent in its impurity surface concentration.

The impurity vapor pressure produced by the source can also be altered by a change in the chemical nature of the source. For example, a change can be produced by loss of residual water vapor at high temperature. The experimental evidence to be given later indicates that such chemical changes may be important in a practical situation.

A certain amount of impurity vapor will be lost by diffusion into the walls of the box. If these walls are exposed to the impurity vapor for a time long enough to make the concentration gradient in the walls small, the flow into the walls will become negligible. A practical procedure is to make one diffusion run for the purpose of saturating the walls of the box.

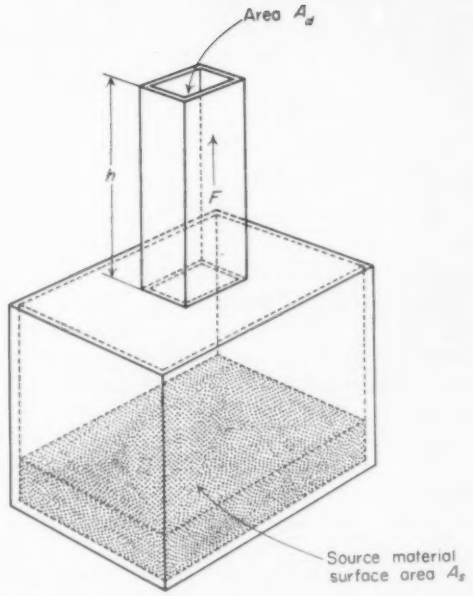


FIG. 4. An idealized box with an opening through a duct of area A_d and height h.

3. EXPERIMENTAL

(A) *Diffusion equipment and procedure*

A schematic diagram of the diffusion equipment is shown in Fig. 5. The equipment consists basically of a box or capsule made of quartz, a quantity of source and the silicon to be diffused.

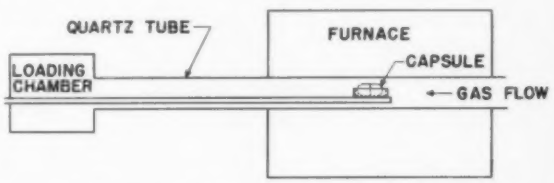


FIG. 5. Diagram of the equipment used for diffusion.

The box containing source and silicon is placed in a quartz tube passing through an electrically heated furnace (Sentry Model V). The temperature of the furnace is controlled by a Wheelco controller (Model 252P) working from a platinum–(platinum, 10 per cent rhodium) thermocouple placed outside the diffusion tube and in the center of the hot zone. A carefully dried gas of controlled

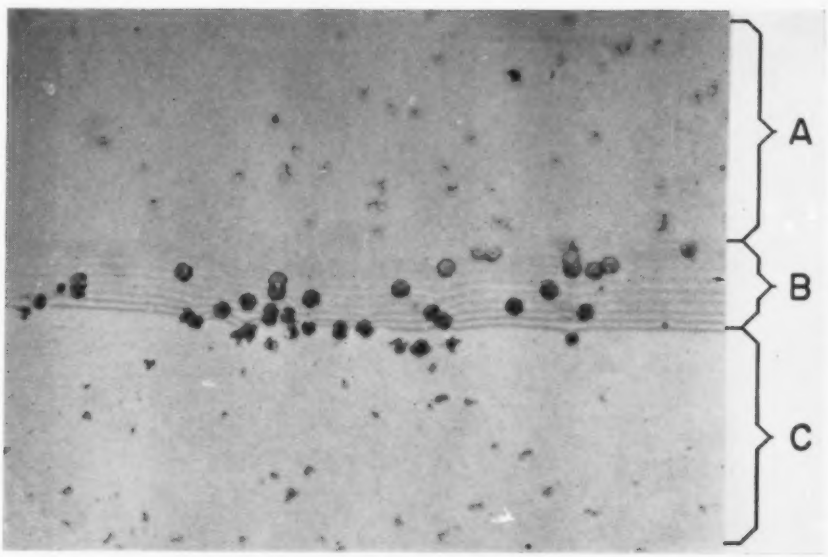


FIG. 6. Amorphous silicon oxide containing small crystals of crystobalite.

Region *A:* Oxidized silicon with crystalline areas faintly visible.
Region *B:* Oxide tapered in thickness. Crystalline areas prominent. Crystal diameters about 5×10^{-3} cm.
Region *C:* Silicon surface free of oxide. Silicon is faintly marked where crystals were present.

[*facing p.* 6

composition and flow rate is introduced at one end of the tube.

The source material is preheated for 10 min at the diffusion temperature. After cooling in the loading chamber, the cleaned silicon samples are placed in the box in the same atmosphere used in the diffusion tube. The box is then placed in the hot zone of the furnace for the desired time of diffusion.

The diffused layers are characterized by measurements of the sheet resistivity and the p–n junction depth. A four-point-probe apparatus with a constant current source and a potentiometer is used to measure the sheet resistivity. This apparatus is capable of measurements with a reproducibility of about 3 per cent. The depths of the diffused layers are measured by the angle-lap and stain technique.[5] The angle is measured with a goniometer. The depth obtained in this way is accurate to about 5 per cent. The surface concentration is calculated from the sheet resistivity and layer depth under the assumption of a complementary error-function distribution, using the relations given by BACKENSTOSS[6].

(B) *Source material*

The purity of the SiO_2 which is used for the diluent in the source has a large influence on the perfection of the SiO_2 layer which is grown on the silicon sample. The appearance of an oxide grown in the presence of a source which contained a fraction of one per cent of sodium is shown in Fig. 6. Crystalline regions of oxide can be seen imbedded in the amorphous oxide layer. The crystals are most readily observed following slight etching with HF. In Fig. 6 a region of oxide with a tapered thickness was formed by dissolving in HF. The crystalline regions were identified by means of electron and X-ray diffraction as being the crystobalite form of quartz.* Other impurities are expected to have a similar effect on the oxide layer. In particular, preliminary observations indicate that the presence of phosphorus results in more numerous crystalline regions than is the case with no impurity oxide deliberately introduced. However, phosphorus does not enhance the growth nearly as much as does sodium. In contrast with phosphorus, boron has little effect.

By using SiO_2 from which sodium (as judged by a flame test) and other impurities were removed by chemical means, the frequency of appearance of the crystalline regions was greatly reduced. An even purer source of SiO_2 was obtained from tetraethylorthosilicate,* (hereafter abbreviated as TEOS). This liquid rapidly hydrolyses in the presence of ammonia to produce a hydrated form of SiO_2 (silicic acid).

The use of TEOS enables one to mix the impurity oxide with the SiO_2 in an intimate fashion. This is done by dissolving the impurity oxide in water (B_2O_3 and P_2O_5 are adequately soluble) and adding ethyl alcohol to make a solution with TEOS, since water is immiscible with TEOS. The hydrolysis is then carried to completion by adding an excess of ammonia. After precipitation of the silicic acid, the ethyl alcohol and water are slowly evaporated to leave a powder. In order to keep the powder dry, it is maintained at 100°C.

The physical arrangement of the impurity in this powder is probably not a true solution, but it is certainly an intimate mechanical mixture, since the particles of SiO_2 have dimensions of only a few hundred angstroms.[7] To produce a solution and to drive off residual water the source material is heated at the diffusion temperature (1200–1300°C) for about 10 min in the closed box, just before adding the silicon to be diffused. The specific surface area of such material as a function of sintering temperature is discussed by VAN NORDSTRAND *et al.*[8] A graph of their results is given in Fig. 7, where an additional point is plotted at the melting point of quartz.

(C) *Effect of oxygen concentration*

The purpose of including oxygen in the atmosphere is to form a protective oxide layer on the silicon. Our experiments indicate that at temperatures up to 1300°C a concentration of 1 per cent O_2 in N_2 is about the lowest oxygen concentration which will prevent pitting of the surface.

The concentration of oxygen in the gas has a strong effect on the surface concentration of the impurity in the silicon. In the case of a 1 per cent phosphorus source (Table 1), a surface concentration of $5{\cdot}5 \times 10^{18}$ atoms/cm^3 was obtained using 21 per cent O_2 in N_2, while with the same source

* Thanks are due to Mrs. E. A. WOOD and Mrs. M. H. READ for help in identifying this form of quartz.

* Obtained from Union Carbide and Carbon Chemicals Corporation.

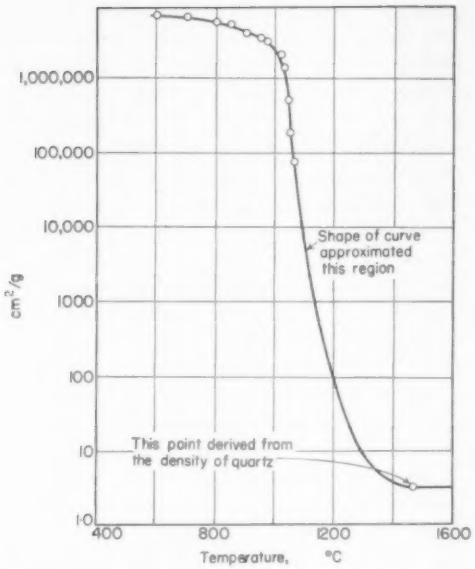


FIG. 7. Specific surface area of silica gel as a function of sintering temperature. Gel held at a given temperature for 12 hr. (After VAN NORDSTRAND *et al.*[8])

using 1 per cent O_2 in N_2, the surface concentration increased to $8{\cdot}1\times10^{19}$ atoms/cm^3. In a typical case of boron diffusion, the surface concentration of boron diffused in pure oxygen was 5×10^{18} atoms/cm^3, while with the same source used in 2 per cent O_2 in N_2, the surface concentration increased to $1{\cdot}8\times10^{20}$ atoms/cm^3.

(D) *Effect of time*

In order to determine the rate of depletion of the source material, a series of diffusion runs was made, always using the same source. A group of four samples was diffused and measured in each run.

A source containing 0·01 per cent B_2O_3 was used at 1300°C. At the end of 1-hr periods the samples were removed and a new set of samples was inserted. Proceeding in this way the concentration of the source was sampled. The surface concentration in the silicon is plotted versus time in Fig. 8. In the Appendix a calculation of the rate of depletion of the surface of this source under a reasonable set of assumptions is given. The calculation indicates that there would be a 10 per cent decrease after 70 hr. The experiment revealed an *increase*

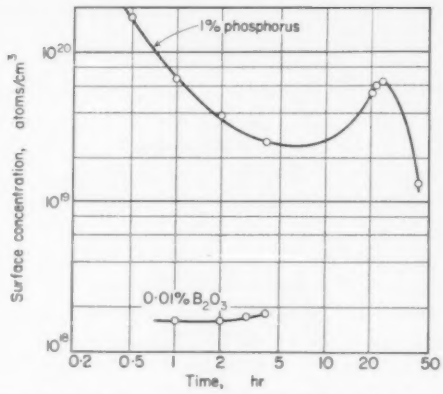


FIG. 8. Surface concentration of diffused layers vs. total time of heating the source material.

of 12 per cent in surface concentration after 4 hr. The increase might be due to loss of some residual water vapor from the source material, which reduced the amount of oxygen available within the capsule.

In a similar experiment a source containing 1 per cent phosphorus was used at 1250°C. At the end of ½-hr periods the samples were removed and new samples inserted. In Fig. 8 the results indicate an initial fall in surface concentration of one order of magnitude in 8 hr, followed by a rise to a maximum at 25 hr. A fall can always be interpreted as a depletion, while a rise might be due to loss of water vapor. A full understanding of this observation will require further investigation.

The experimental observations on the effect of time indicate that for the best reproducibility using boron or phosphorus a fresh source should be used for each diffusion run. That practice was followed in obtaining the data discussed in the next section.

(E) *Diffusions under standardized conditions*

A summary of the source compositions, conditions of diffusion and results which were obtained using sources containing B_2O_3 and P_2O_5 are given in Table 1. For these diffusions the gas composition was maintained at 1 per cent O_2–99 per cent N_2, flowing at 1500 cm^3/min.

The experimental results for each source composition were obtained from four or more runs, with three or more samples in each run. The parameters measured within a run were characterized

Table 1. Source compositions, conditions of diffusion and results obtained using sources containing B_2O_3 *and* P_2O_5. *Gas composition 1 per cent* O_2: *99 per cent* N_2

Source composition by weight	Parent Si $\lvert N_d - N_a \rvert$ (atoms/cm^3)	Temp. (°C)	Time (min)	No. of runs	Samples per run	Sheet resistance		Junction depth		C_0 (atoms/cm^3)
						Value (Ω)	3 σ limits (%)	Value (mil)	3 σ limits (%)	
10% B_2O_3:1·5% B	2×10^{16}	1300	45	4	3	2·18	±4	0·32	± 3	$1{\cdot}8\times10^{20}$
0·01% B_2O_3:0·0015% B	$4{\cdot}5\times10^{16}$	1300	60	4	4	607	±17	0·12	± 8	$1{\cdot}5\times10^{18}$
5% P:1·5% P_2O_5	6×10^{16}	1300	60	4	4	0·713	±23	0·37	±16	$2{\cdot}4\times10^{20}$
1% P:2·3% P_2O_5	2×10^{17}	1250	60	11	4	3·7	±14	0·20	±40	$8{\cdot}1\times10^{19}$
1% P:23% P_2O_5	3×10^{16}	1250	60	4	4	277	±37	0·069	±52	$1{\cdot}4\times1018$

by a mid-range and a range. The median of the mid-ranges was taken to represent the experimental point. A statistical analysis using these mid-ranges and ranges indicates that the process is in statistical control. The 3σ control limits are used to indicate the controllability of the process.

Each composition was tested for the masking produced by an oxide layer on 0·3–0·5 Ω-cm silicon samples. A standard oxide layer 5000 Å thick was formed by exposing the silicon samples to steam at a temperature of 1300°C for 30 min. The oxide was then removed in stripes by the use of HF and a protective wax coating. These striped samples were included in the diffusion runs, together with unoxidized samples. The striped samples were angle-lapped and the region near the oxide boundary was examined to determine the effectiveness of the oxide masking. All the sources which were used here produced oxide masking under the conditions reported in Table 1. A photograph of an angle-lapped and stained specimen is shown in Fig. 9. Region *A* is the surface of the sample and shows oxidized areas and an oxide-free stripe. Region *B* is the angle-lapped and stained surface[5] showing a junction 0·25 mil below the surface of the sample.

When pure B_2O_3 was used as a source a new phase formed on the surface of the silicon. This material appeared as a gray or black deposit. It is almost certainly a boron–silicon compound. This phase can be eliminated by using a source containing sufficiently diluted B_2O_3.

(F) *Simultaneous diffusions through an oxide layer*

Preliminary investigations of masking indicate that boron penetrates an oxide layer more rapidly than does phosphorus. This difference was demonstrated as follows.* Inasmuch as the diffusion constants for boron and phosphorus in elemental silicon are nearly identical, the difference in their rate of penetration through an oxide layer can be utilized to make an *n–p–n* structure in *n*-type silicon. Such a structure will result, provided that the phosphorus is sufficiently concentrated so that it will produce a higher surface concentration in the silicon than will the boron, and also provided that the boron penetrates the intervening oxide layer before the phosphorus does.

* This experiment was carried out by M. O. THURSTON.

A source material was used which contained more phosphorus (0·5 per cent) than boron (0·1 per cent). In order to observe the effect of various oxide thicknesses, a graded oxide was prepared from an initially uniform layer by etching in HF while slowly withdrawing the sample from the solution. The resulting oxide thickness varied from 2000 Å to zero. The parent n-type silicon had a net donor concentration of 10^{16} atoms/cm^3. The diffusion was carried out for 1 hr at 1250°C in an atmosphere of 21 per cent O_2 in N_2. The sample was then angle-lapped at right-angles to the direction of oxide thickness gradation. The junctions are shown in Fig. 10. That portion of the sample which had the thickest oxide is at the right. There the silicon shows only a p-type layer (region B) formed by the boron diffusion. The portion of the sample which had initially zero oxide is at the left and shows no junctions since the phosphorus predominates over the boron. Between these extremes an n–p–n structure is seen. The line ab marks the location where the initial oxide thickness was 1200 Å.

This experiment can be used to obtain estimates of the effective diffusion constants for phosphorus and boron in SiO_2. The phosphorus barely penetrates the SiO_2 at a thickness near 2000 Å. On unoxidized silicon the phosphorus would have produced a surface concentration of $\sim 10^{20}$ atoms/cm^3 in the silicon. In the present case the surface concentration of the phosphorus in the silicon is estimated to be about 10^{-2} times the value in the unoxidized case. The ratio of the concentration of the phosphorus in the SiO_2 2000 Å below the surface to the concentration at the surface is therefore also expected to be 10^{-2}. Using the solution to the diffusion equation for a constant source concentration

$$x = 2(Dt)^{1/2}\,\mathrm{erfc}^{-1}(C/C_0) \qquad (5)$$

one obtains $D \sim 10^{-14}$ cm^2/sec for phosphorus in SiO_2 at 1250°C. This may be compared with the case of phosphorus in silicon at 1250°C, where $D = 7\times10^{-12}$ cm^2/sec. Owing to the slow dependence of $\mathrm{erfc}^{-1}(C/C_0)$ on C_0, the effective diffusion constant for boron in SiO_2 is unlikely to be more than twice that for phosphorus.

4. CONCLUSIONS

Diffusion in a loosely closed box is suitable for producing p- or n-type diffused layers of controlled area (that is, masked) with surface concentrations as high as 2×10^{20} atoms/cm^3. The system contains only those chemical elements which are most essential, namely silicon, impurity and oxygen, and nitrogen as a diluent. Other elements besides boron and phosphorus could also be diffused and masked by this method.

Several aspects of the processes at the surface of the silicon need further clarification. The mechanisms of reduction of the impurity oxide and of oxidation of the silicon have a strong influence on the concentration of the impurity in the silicon, but are insufficiently understood. The process of diffusion of impurity oxide through the silicon oxide also needs more study.

In its present state the method provides a flexible means of producing masked diffusions of boron and phosphorus in silicon.

Acknowledgements—This program of investigation of gaseous diffusion in a closed system was initiated by I. M. Ross. Contributions were made to the work by M. O. Thurston, V. Fong and C. E. Golightly. I thank F. M. Smits for continued guidance.

APPENDIX

As was mentioned above, in order to analyze the depletion of the source, three factors should be considered:

(1) The diffusion of impurities out of the source.
(2) A possible rate process at the surface of the source.
(3) The pumping speed of the duct.

The diffusion out of the source can be described by a diffusion coefficient D. In reality, D may be a strong function of concentration.

A rate process at the surface of the source can be described by the rate constant K, which relates the flow F across the surface to the difference between the actual surface concentration and the equilibrium concentration corresponding to the density in the gaseous phase:

$$F = [N(0,t) - N_g k_g]K \qquad (6)$$

The "gaseous distribution" coefficient k_g in this equation is defined as

$$k_g = N_e/N_g \qquad (7)$$

with N_e the equilibrium concentration in the solid corresponding to the density N_g in the gas phase.

The pumping speed L of the duct relates the flow out of the box to the density in the box (with zero impurity vapor pressure outside)

$$F = LN_g. \qquad (8)$$

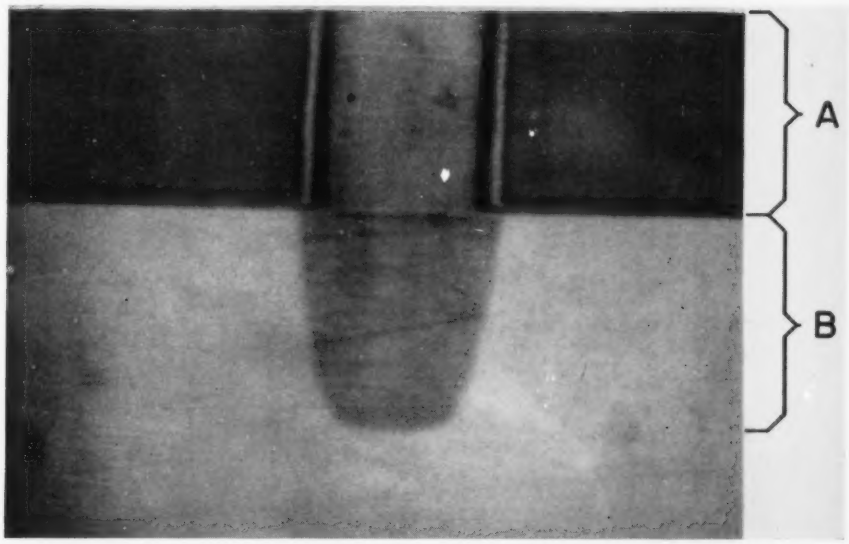


FIG. 9. Masked diffusion of boron with surface concentration of $1{\cdot}8 \times 10^{20}$ atoms/cm^3.

Region *A*: Surface of sample, showing oxidized areas and oxide-free stripe.
Region *B*: Angle-lapped and stained surface, showing junction 0·25 mil below the surface of the sample.

[*facing p.* 10

FIG. 10. *n–p–n* structure formed by simultaneous diffusion of boron and phosphorus through an oxide layer.

Region *A*: Angle-lapped and stained surface.
Region *B*: *p*-Type region 0·3 mil thick.
Region *C*: Surface of sample from which oxide was removed during staining.
Region *D*: Oxidized surface of sample.

With these definitions, the net flow of impurity out of the source can be described by a generalized rate constant as[9]

$$F = N(0, t)K' \tag{9}$$

with

$$K' = \frac{L}{L + k_g K A_s} K \tag{10}$$

in which A_s is the area of the source. Equation (9) serves as a boundary condition for the problem of diffusion out of the source. One can distinguish two extreme cases.

In the case where $L/k_g A_s \to \infty$ the rate limitation is at the surface, since $K' \to K$. Then $F \to N(0, t)K$, and by comparison with equation (6) one sees that $N_g \to 0$.

In the case where $L/k_g A_s \to 0$, the rate limitation is in the duct and

$$K' \to L/k_g A_s \tag{11}$$

where D_g is the diffusion constant for interdiffusion of the impurity vapor and the gas. The values of D_g can be obtained from kinetic theory.[10] From equations (11) and (13), in the case of a rate limitation in the duct,

$$K' = A_d D_g / A_s k_g h \tag{14}$$

The concentration at the surface of the source as a function of time is obtained from a solution of the diffusion equation, with equation (9) as a boundary condition. If the initial uniform concentration in the source is N_0, then[9]

$$\frac{N(0, t)}{N_0} = [\operatorname{erfc} K' \sqrt{(t/D)}] \exp\left(\frac{K'^2 t}{D}\right) \tag{15}$$

This equation is plotted in Fig. 11. In order to maintain the concentration of impurity vapor as constant as possible throughout the time (t) in which a sample of silicon would be exposed for diffusion, one sees from Fig. 11

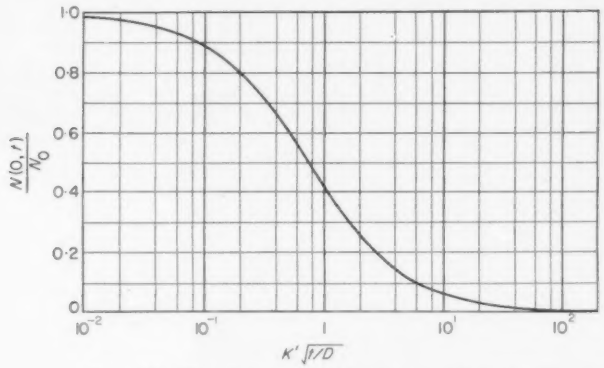


FIG. 11. Fraction of initial concentration in source material $[N(0, t)/N_0]$ vs. $K'(t/D)^{1/2}$.

which does not contain K. This case corresponds to $F \to 0$. From equation (6) one sees that now

$$N_g k_g \to N(0, t) \tag{12}$$

In other words, the vapor pressure within the box will be determined by k_g and the impurity concentration at the surface of the source. This is the case of practical interest. It is approached experimentally by making L small and A_s large. The value of k_g can be estimated from the known vapor pressure of the pure impurity oxide, assuming that vapor pressure decreases linearly with dilution.

The evaluation of L can be treated as a diffusion problem, since the dimensions of the duct are large compared with a mean free path. Furthermore, since the cross-sectional area of the duct is small compared with the cross-sectional area of the box, the gradient of impurity concentration within the box is small compared with the gradient in the duct. One then obtains for L

$$L = A_d D_g / h \tag{13}$$

that it is desirable to make $K'(t/D)^{1/2}$ as small as possible. This can be accomplished by maximizing A_s the source area, minimizing L the pumping speed, and maximizing D the diffusion constant in the source material. Since, in the glassy materials used here as sources, D is known to increase with increasing concentration of impurity, it is to be expected that the higher concentration source materials will maintain a constant impurity vapor concentration for a longer time than will some lower concentration sources. For lower concentration sources, A_s can be increased by finely dividing the material. In the experimental work, this division was accomplished by using SiO_2 precipitated out of a solution.

In the following, a typical calculation is carried out to find the time required to deplete the surface concentration of the impurity oxide by 10 per cent. In order to make a comparison with the experimental results for boron diffusion shown in Fig. 8, we will assume a low concentration B_2O_3 source at a temperature of 1300°C. The capsule dimensions of Fig. 3 will be used.

From Fig. 11, for $N(0, t)/N_0 = 0{\cdot}9$, the value of

$K'(t/D)^{1/2}$ is 10^{-1}, so $t = 10^{-2}D/K'^2$. In the preceding Section F, the diffusion coefficient in the SiO_2 is estimated to be $D = 10^{-14}$ cm²/sec. The rate constant K' is given by equation (14) assuming that the rate limitation is in the duct. Assuming that the vapor pressure decreases linearly with dilution, $k_g(= N_e/N_g)$ is independent of dilution. The vapor density[11] of pure B_2O_3 at 1300°C is 5×10^{14} atoms/cm³. The concentration of pure solid B_2O_3 is 2×10^{22} atoms/cm³, so $k_g = 4 \times 10^7$. From the dimensions of the capsule shown in Fig. 3, $A_d = 2\pi rd \approx 6 \times 10^{-3}$ cm² and $h = 1$ cm. The gaseous diffusion constant D_g for the interdiffusion of B_2O_3 vapor and N_2 at 1300°C can be calculated from kinetic theory[10] to be $D_g = 3$ cm²/sec. The value of A_s can be estimated from the data of VAN NORDSTRAND *et al.*[8] as plotted in Fig. 7. Using a specific area of 20 cm²/g gives $A_s = 20$ cm² for a typical 1-g source, and then $K' = 2 \times 10^{-11}$ cm/sec.

The values for D and K' given above can now be substituted in the expression for t. One then comes to the conclusion that the low-concentration B_2O_3 source will be depleted by 10 per cent in its impurity surface concentration only after 70 hr at 1300°C.

REFERENCES

1. For review articles see C. S. FULLER and C. J. FROSCH, *Transistor Technology*, Vol. 3. Van Nostrand, New York (1958); F. M. SMITS, *Proc. Inst. Radio Engrs, N.Y.* **46,** 1049 (1958).
2. C. J. FROSCH and L. DERICK, *J. Electrochem. Soc.* **104,** 547 (1957).
3. J. P. COUGHLIN, *Heats and Free Energies of Formation of Inorganic Oxides*. Bulletin 542, Bureau of Mines, U.S. Government Printing Office (1954).
4. F. P. HALL and H. INSLEY, *Phase Diagrams for Ceramicists*. American Ceramic Society, Columbus, Ohio (1947).
5. C. S. FULLER and J. A. DITZENBERGER, *J. Appl. Phys.* **25,** 1439 (1954).
6. G. BACKENSTOSS, *Bell Syst. Tech. J.* **37,** 699 (1958).
7. R. K. ILER, *The Colloid Chemistry of Silica and Silicates*. Cornell University Press (1955).
8. R. A. VAN NORDSTRAND, W. E. KREGER and H. E. RIES Jr., *J. Phys. Chem.* **55,** 641 (1951).
9. F. M. SMITS and R. C. MILLER, *Phys. Rev.* **104,** 1242 (1956); R. C. MILLER and F. M. SMITS, *Phys. Rev.* **107,** 65 (1957).
10. S. DUSHMAN, *Scientific Foundations of Vacuum Technique*, p. 75. Wiley, New York (1949).
11. J. R. SOULEN, P. STHAPITANONDA and J. L. MARGRAVE, *J. Phys. Chem.* **59,** 132 (1955).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 13–21. Printed in Great Britain

A NEW DEVICE USING THE TUNNELING PROCESS IN NARROW *p–n* JUNCTIONS

LEO ESAKI and YURIKO MIYAHARA

Sony Corporation, Tokyo, Japan

(*Received* 8 *June* 1959; *in revised form* 24 *September* 1959)

Abstract—The design, fabrication and electrical characteristics of a new device made of heavily doped silicon or germanuim for use as an active circuit element, are discussed. The device is essentially a single, narrow *p–n* junction made by alloying techniques, and the d.c. negative resistance in its characteristics arises from a tunneling process across the junction, which makes the device inherently capable of working at very high frequency. The device, which is simple in construction and stable even in high humidity, is promising for low-level, high-speed switching, oscillation and amplification.

Résumé—Le dimensionnement et la fabrication et les caractéristiques électriques d'un nouvel appareil utilisant du germanium ou du silicium très dopé et fonctionnant comme un élément actif sont décrits dans l'article. L'appareil est essentiellement une jonction *p–n* unique très étroite obtenue par des méthodes d'alliage et la résistance négative qui apparaît dans ces caractéristiques résulte d'un effet tunnel à travers la jonction, ce qui rend l'élément capable de fonctionner à de très hautes fréquences. L'appareil est de construction simple, stable même en atmosphère très humide et semble très intéressant pour la commutation, l'amplification ou la production d'oscillations de faible puissance.

Zusammenfassung—Es werden Anordnung, Herstellung und elektrische Eigenschaften eines neuen Bauelementes besprochen, das aus stark dotiertem Silizium oder Germanium zum Zwecke der Anwendung als aktives Schaltelement hergestellt wurde. Das Bauelement ist ein durch die Legierungstechnik gewonnener enger *p–n*-Übergang; der negative Widerstand in seiner Gleichstrom-Kennlinie rührt von dem Tunnelprozess durch den *p–n*-Übergang her, der das Bauelement in den Stand setzt, bei sehr hoher Frequenz zu arbeiten. Das Bauelement, das einfach im Aufbau und selbst bei hoher Feuchtigkeit beständig ist, verspricht, für Niederspannungsschaltung bei hoher Geschwindigkeit, Schwingungserzeugung und Verstärkung interessant zu werden.

1. INTRODUCTION

THE interesting properties revealed in extremely narrow germanium *p–n* junctions were previously reported by ESAKI[1,2]. In the subsequent study, we have succeeded also in the observation of negative resistance and inverted rectification in extremely narrow silicon *p–n* junctions. Although the general behaviour is similar to that of germanium, the range of the negative resistance can be much extended, as has been expected. This new phenomenon, due to tunneling of the charge carrier in these single *p–n* junctions of germanium and silicon, is not only theoretically interesting because of an elementary process in the physics, but also quite promising for obtaining active semi-conductor circuit elements at high frequencies. The name "Tunnel Diode" or "Esaki Type Diode"[3] has been given to this device.

It is the purpose of the present paper to describe the process of fabrication and electrical characteristics of this new device and to discuss it.

2. PROCESS OF FABRICATION

Heavily doped silicon single crystals which made narrow junctions possible were pulled in the (111) direction with the usual Czochralski method, using phosphorus as doping impurity.

The donor (phosphorus) and the electron concentrations of these crystals were determined by colorimetric analysis and the Hall coefficient and resistivity measurements, respectively. In Table 1, the donor and electron concentrations of several

typical crystals are shown with the specific resistivity and the electron mobility. These crystals have no indication of inhomogeneous distribution of phosphorus in spite of the high impurity concentration of more than 10^{20} cm^{-3}. The experimental results indicate that the percentage of ionized donor impurity at room temperature decreases gradually from nearly 100 per cent to 60 per cent with increase from $0{\cdot}68\times10^{20}$ cm^{-3} to $2{\cdot}6\times10^{20}$ cm^{-3} in impurity concentration. Abrupt and narrow silicon p–n junctions were satisfactorily fabricated by alloying techniques using aluminium containing boron. Junctions of the area over the range from $1{\cdot}2\times10^{-3}$ cm^2 to 3×10^{-4} cm^2 were made. A gold wire of approximately 100 μ diameter was attached to the aluminium dot with the usual thermo-compression bonding as shown in Fig. 1(a) and the diode was assembled to a rugged construction similar to commercial microwave silicon diodes, as shown in Fig. 1(b), giving a series resistance of the order of 0·1 Ω.

3. ELECTRICAL CHARACTERISTICS

Current–voltage characteristics of a typical diode made with silicon of $1{\cdot}1\times10^{20}$ cm^{-3} phosphorus concentration are illustrated in Fig. 2 and Fig. 3, as compared with the case of germanium. Fig. 2 shows the characteristics over the low-voltage range of the "forward" and "reverse" directions at room temperature. The present p–n junctions are more conductive in the reverse direction than in the forward direction. Fig. 3 shows also the semi-log plot of the characteristics at 353°K, 298°K, 195°K, 77°K and 4·2°K in the forward direction. Over the voltage range higher than 0·7 V in the forward direction, the current–voltage curve can be fitted almost quantitatively by the well-known relation:

$$I = I_s(T)\{\exp(qV/kT)-1\}.$$

The numerical values of $I_s(T)$ are not much different from the usual silicon and germanium p–n junctions and following approximately the relation:

$$I_s(T) = I_0 \exp\left(\frac{-q\epsilon_G}{kT}\right),$$

where ϵ_G is the energy gap. The diode theory will be applicable for the present junction because the mean free path of the carriers is comparable with the p–n junction width.

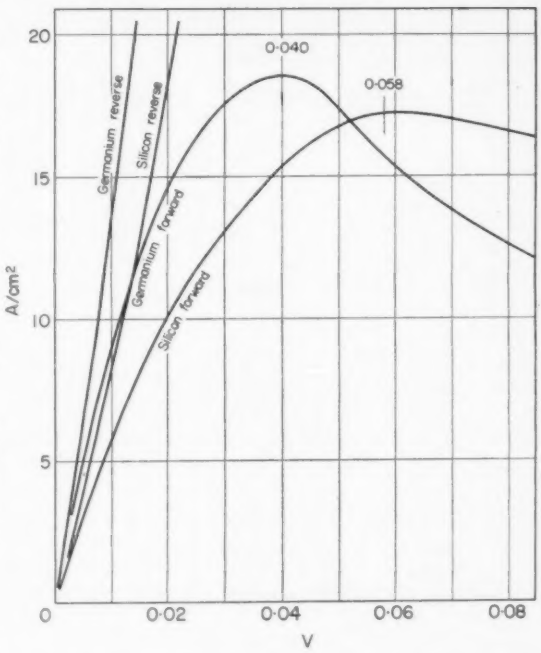


FIG. 2. Current–voltage characteristics of silicon and germanium over the low-voltage range in the reverse direction and in the forward direction at room temperature.

Considerable tunneling current is observed at lower voltage before the normal diode current becomes dominant[2,4] and the maximum and the minimum of the curve are observed approximately at 0·06 V and 0·35 V, respectively, in every specimen, as shown in Fig. 2 and Fig. 3. The ratio I_{max}/I_{min} is from 2 to 4.

Fig. 4(a), (b) and (c) shows current–voltage characteristics of specimens made with silicon of 7×10^{19} cm^{-3} and $1{\cdot}2\times10^{20}$ cm^{-3} and with germanium of $1{\cdot}9\times10^{19}$ cm^{-3} phosphorus concentration, respectively, over the low forward voltage range at 303°K, 90°K and 4·2°K. While the current increases with decrease in temperature in the case of germanium, Fig. 4(c), the silicon specimens of Fig. 4(a) and (b) show an opposite tendency with decreasing temperature. Moreover, it should be noted that the curve for silicon at 4·2°K reveals definitely some fine structure and has two bending points A and B. The curve is considered to consist of at least three components. This result will be

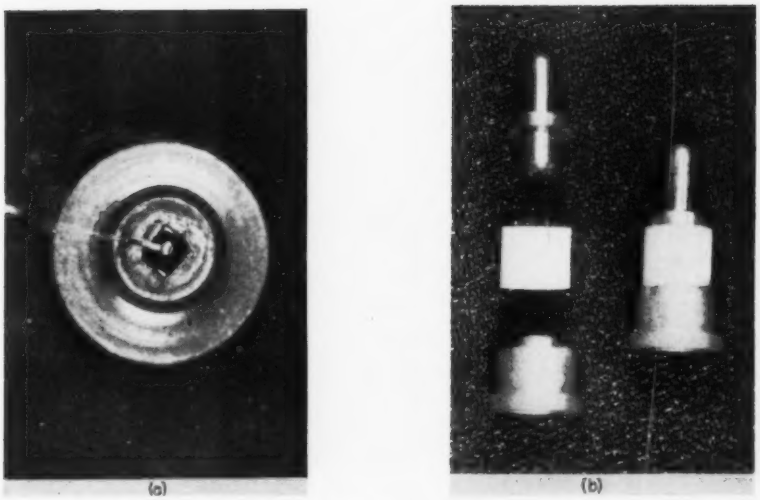


FIG. 1 (a) and (b). The diode and its construction.

[*facing p.* 14

explained by introducing the concept of phonons associated with a tunneling process, which will be discussed in detail in another paper.

The capacitance C measured with a r.f. bridge

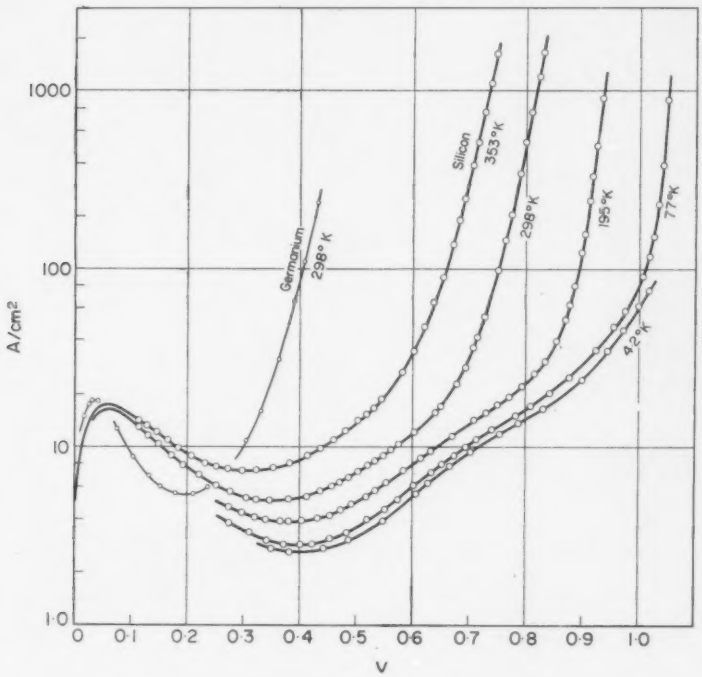


FIG. 3. Current–voltage characteristics in the forward direction at various temperatures.

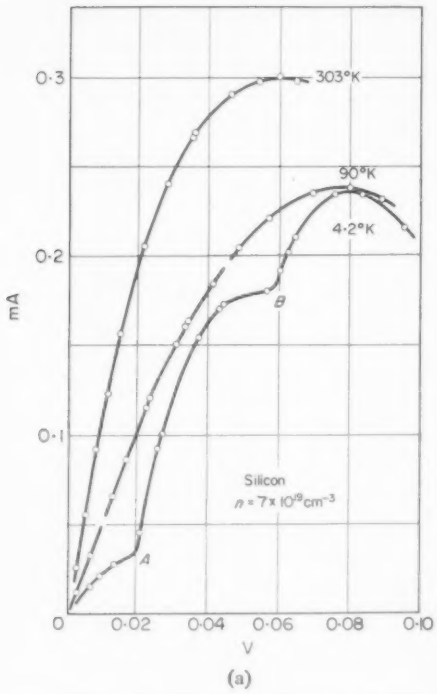
is plotted against bias voltage in Fig. 5. The measured points are in agreement with the familiar relation in the abrupt p–n junction between the capacitance C and the applied voltage V:

$$C = \frac{k}{\sqrt{(\epsilon_G - V)}}$$

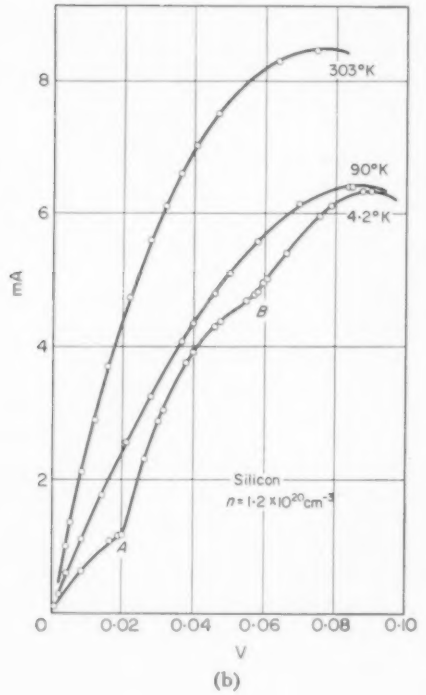
where k is constant. The dotted line in Fig. 5 shows the positive and negative R in the equivalent circuit of Fig. 6.

These values of C and R depend only upon the donor concentration in the n-type silicon up to 10^{20} cm^{-3}. However, at concentrations greater than this value, they become sensitive to the alloying temperature and time, and the content of boron in the aluminium dot. This fact indicates that it would be rather difficult to make the excess acceptor concentration in the regrowth region increase up to more than 10^{20} cm^{-3} by alloying to the n-type base silicon containing phosphorus of more than 10^{20} cm^{-3} in concentration.

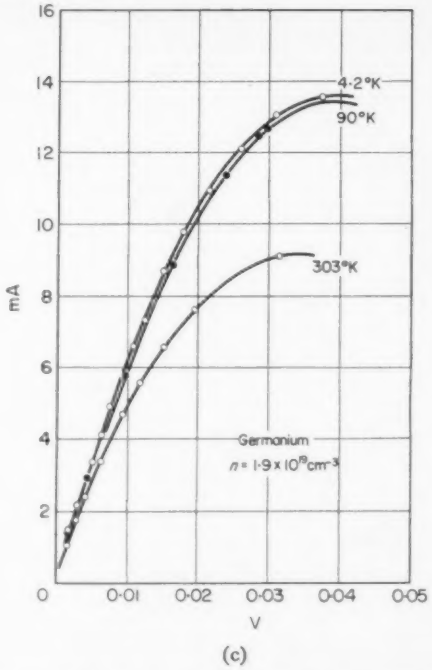
The effect of pressure on the electrical characteristics should also be mentioned. When the p–n junctions were pressed perpendicularly with a needle having a radius of 25 μ or more at the top, the electrical characteristics were changed reversibly, as shown in Fig. 7(a), (b) and (c), according to the pressure strength. The tunneling current increases linearly with increase of the pressure. Fig. 8 shows a change of $(dI/dV)_{V=0}$, I_{max} and I_{min} with the increase of the pressure. The applied pressure under the contact between the needle and the aluminium dot is estimated as approximately 10^7 g/cm^2 at scale 5 of the unit shown in Fig. 8. The increase of the tunneling current observed in



(a)



(b)



(c)

FIG. 4 (a), (b) and (c). Current–voltage characteristics of specimens made with (a) silicon of 7×10^{19} cm^{-3}, and (b) silicon of $1{\cdot}2 \times 10^{20}$ cm^{-3} in phosphorus concentration and (c) germanium of $1{\cdot}9 \times 10^{19}$ cm^{-3} in phosphorus concentration over the low forward-voltage range at 303°K, 90°K and 4·2°K.

this experiment is much larger than expected from the effect that the p–n junction is made narrower owing to the elasticity of silicon.

4. DISCUSSION

The current–voltage characteristics and their temperature dependency are explained with the same tunneling mechanism as described in the previous paper.[1,2] That is, the negative resistance arises from a tunneling of electrons from filled levels in the conduction band and the impurity band on the n_+-side of the junction through the narrow barrier to empty levels of equivalent energy in the p_+-side as forward bias is applied. When using silicon, we find that the range of the negative resistance can be much extended owing to the larger energy gap than that of germanium, while the general shape of the curve becomes broader.

The tunneling probability of the electrons

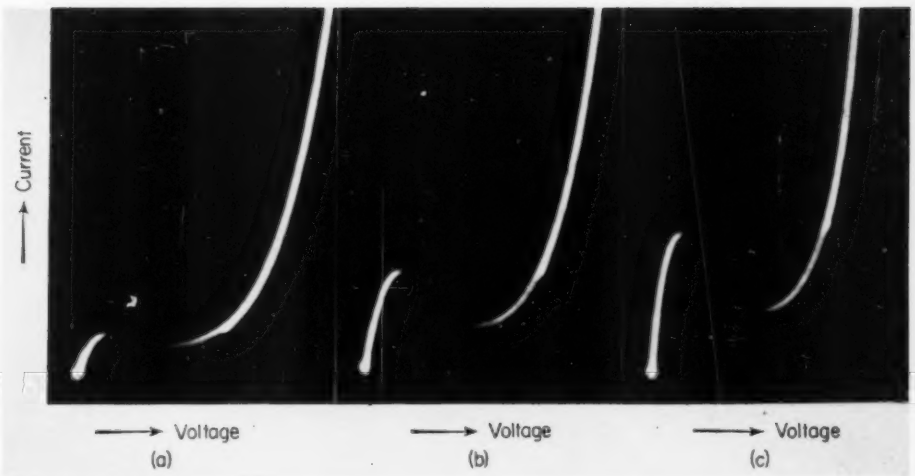


FIG. 7. The pressure effect on the current–voltage characteristics. The characteristics (a) are reversibly changed to (b) and (c) with the increase in the applied pressure.

[facing p. 16

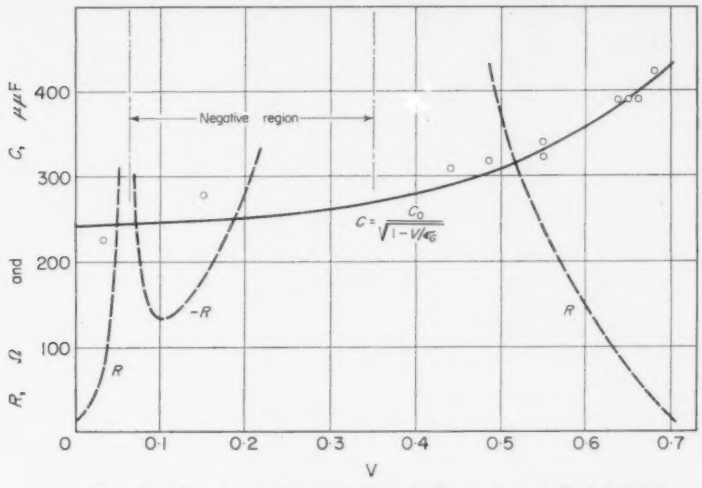


FIG. 5. The dependency of R and C on the applied voltage.

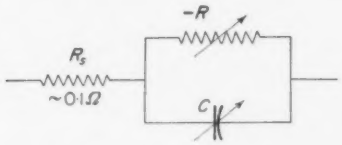


FIG. 6. The equivalent circuit of the device.

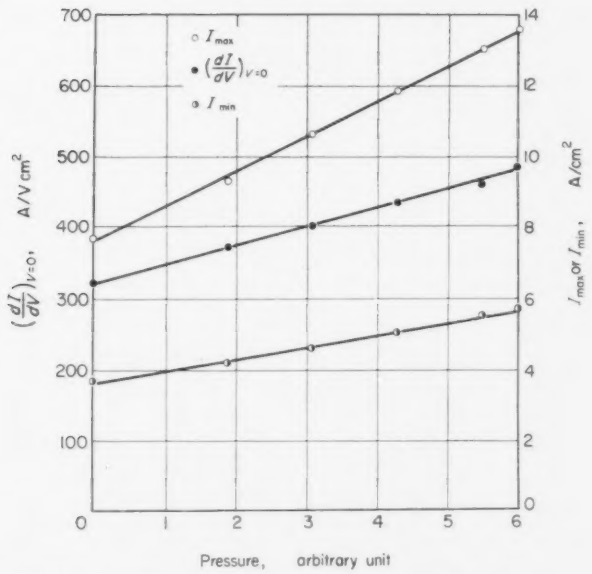


FIG. 8. $(dI/dV)_{V=0}$, I_{max} and I_{min} versus applied pressure.

through the energy gap, due to the built-in field in the p–n junction, is roughly given by[5]

$$Z = \exp\left(-\frac{\pi^2 m^* a \epsilon_G^2}{h^2 q F}\right) \qquad (1)$$

where m^*, a, ϵ_G, h, q and F are effective mass of electrons, lattice constant, energy gap, Planck's constant, electronic charge and field strength, respectively. The above formula is modified for larger energy gaps as follows,[6]

$$Z = \exp\left(-\frac{\pi^2 (2m^*)^{1/2} \epsilon_G^{3/2}}{hqF}\right) \qquad (2)$$

The field F can be estimated from the formula

$$F = \frac{4\pi C \epsilon_G}{\kappa}$$

with the measurement of the capacitance C, where κ is the dielectric constant. The relative value of the tunneling probability would be derived from the tangent (dI/dV) at $V = 0$ in the current–voltage curve. Fig. 9 shows $\log(dI/dV)_{V=0}$ versus $1/F$ obtained with many silicon and germanium specimens. Two straight lines are obtained for silicon and germanium, respectively, though the measured points are scattered because of a considerable patch effect.† The slopes give the following numerical values for the tunneling probability,

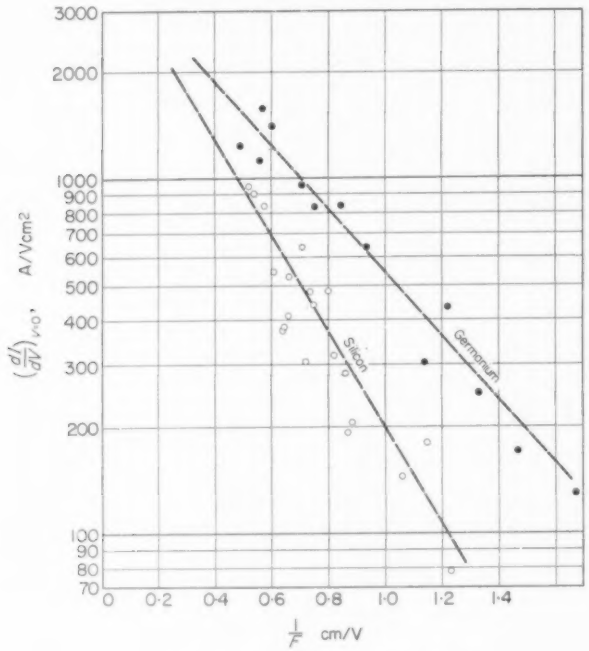


FIG. 9. $\text{Log}(dI/dV)_{V=0}$ versus $1/F$.

$$Z = \exp\left(-\frac{2\times 10^6}{F}\right) \text{ in germanium and}$$
$$Z = \exp\left(-\frac{3\times 10^6}{F}\right) \text{ in silicon} \qquad (3)$$

Using formula (1), the effective mass m^* can be calculated as $m^* = 0{\cdot}4m$ in germanium and $m^* = 0{\cdot}25m$ in silicon. These will indicate average values of longitudinal and transverse masses because the junction surface is not an exact (111) face.

† Fig. 9 and equation (3) may be revised by measuring more uniform junctions in future.

It is plausible that these differ from the values determined by more exact measurement[7] for purer specimens than our degenerate silicon and germanium.

Further, it should be noted that the density-of-states masses for both the conduction band and the valence band in silicon are twice as much as those in germanium so that the state densities for those bands are roughly 2·8 times more than in germanium. Therefore, in order to obtain the same position of the Fermi level, the carrier concentration in silicon should be 2·8 times more than that in germanium. This tendency coincides with our experimental results that more heavily doped silicon crystals than germanium were needed, as shown in Table 1.

A low effective electron mass in the intermetallic semiconductors is most attractive for this new phenomenon and device. In order to extend the negative-resistance region, a large energy-gap intermetallic semiconductor may be used as the material. It is quite favourable that the starting material of an intermetallic semiconductor is not required to be extra pure.

Table 1

	Phosphorus concentration ($\times 10^{19}$ cm^{-3})	Electron concentration ($\times 10^{19}$ cm^{-3})	Specific resistivity ($\times 10^{-4}$ Ω-cm)	Electron mobility (cm^2 V^{-1} sec^{-1})
silicon	6·8	6·7	10·9	85
silicon	11·0	9·9	7·8	81
silicon	11·0	9·8	7·5	85
silicon	16·0	12·0	6·7	78
silicon*	26·0	16·0	5·5	71
germanium	1·6	1·6	9·1	426
germanium	3·4	3·4	6·7	268

* This specimen was not used to make the diode.

5. CONSIDERATION FOR APPLICATIONS

The assembled diode is highly reliable because the construction is simple and the metal-like surface of the degenerate semiconductor is quite stable even in a highly humid atmosphere. This device is of use for such active circuit elements as amplifiers or oscillators which are inherently capable of working at very high frequency.[8]

The maximum oscillation frequency f_{max} is simply derived as follows:

The criterion for oscillation is

$$|\text{negative } R| \leqslant \frac{L}{R_s C}.$$

Under the threshold condition, where |negative R| is equal to $L_0/R_s C_0$, f_{max} is given by

$$f_{max} = \frac{1}{2\pi\sqrt{(L_0 C_0)}} = \frac{1}{2\pi C_0\sqrt{(R_s|-R|)}} \quad (4)$$

In the case of the specimen shown in Fig. 5, fabricated with silicon of 10^{20} cm^{-3} electron concentration,

$$R_s = 0{\cdot}1\ \Omega$$
$$|-R| = 140\ \Omega \text{ and}$$
$$C_0 = 200\ \mu\mu f,$$

so that f_{max} is calculated as 200 Mc/s.

Fig. 10 shows a simplified equivalent circuit for amplifier. I and g_1 are a current source and a source conductance, respectively. The device is shown with a parallel arrangement of a capacitance C_g and

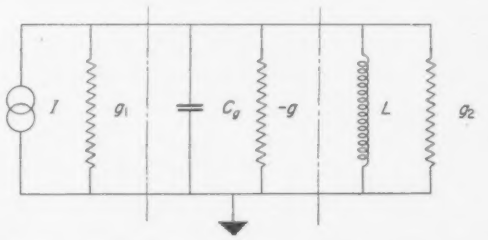


FIG 10. The simplified equivalent circuit for amplifier.

a negative conductance $-g$. L and g_2 are an inductance coupled with C_g and a load conductance, respectively. We can simply derive the power gain P_G, the current gain I_G and the bandwidth B_W, if the impedance matching is provided by $g_1 = g_2 - g$, as follows:

$$P_G = \frac{4g_1g_2}{(g_1+g_2-g)^2} = \frac{g_2}{g_1},$$

$$I_G = \frac{g_2}{g_1} \qquad \text{and}$$

$$B_W = \frac{(g_1+g_2-g)}{2\pi C_g} = \frac{g_1}{\pi C_g}.$$

Therefore the gain-bandwidth product is

$$I_G \cdot B_W = \frac{g_2}{\pi C_g}.$$

In the case of limiting condition, $g_1 \to 0$ and $g_2 \to g$, the gain-bandwidth product becomes

$$I_G \cdot B_W = \frac{g}{\pi C_g}.$$

If the equivalent circuit of the device in Fig. 6 is transformed to the parallel arrangement of C_g and $-g$ shown in Fig. 10, we can obtain the next formula,

$$I_G \cdot B_W = \frac{1}{\pi RC}\left\{1-\left(\frac{f}{f_{\max}}\right)^2\right\} \qquad (5)$$

The gain-bandwidth product of our above-described silicon specimen, of which the frequency response is shown in Fig. 11, is 11 Mc/s over the low frequency range.

It is a natural consequence that $f_{\max}$ and the gain-bandwidth product in our two-terminal device is independent of the junction area because the product $C \cdot \sqrt{(R_s R)}$ in formula (4) and the product RC in formula (5) are invariable to the area. It follows also that the junction area should be made small in the case of such three-terminal devices as transistors for high frequency use because of the sheet resistivity in the base region, a fact which makes high power and also high frequency operation in the three-terminal device impossible.

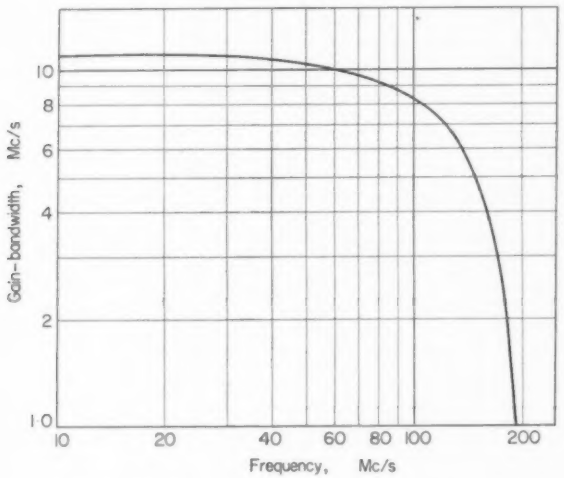


FIG. 11. The frequency response of the gain-bandwidth.

It is of course necessary to reduce R_s in order to improve the high-frequency response of the present device. Besides, it is required to decrease the junction width by increasing the carrier concentration. If the junction width is decreased to approximately one-half, the value $|-R|$, which is inversely proportional to the value Z in the formula (3), should drop to a fraction of the original value, while the capacitance is only doubled. This should

result in increasing both f_{max} and the gain-bandwidth product. The charge carrier concentration in germanium crystals used in the present study is much less than in silicon crystals, as indicated in Table 1.[2] If the device is made with a germanium crystal of near 10^{20} cm^{-3} carrier concentration, it would be possible to obtain better frequency response and larger gain-bandwidth product than that shown in Fig. 11 without any difficult technical problems. Another method to make possible high-frequency operation is to use a compound semiconductor of a low electron effective mass such as InSb, InAs or GaAs, as is described in the preceding paragraph.

The n-shaped characteristics of the device are quite favourable for pulse generator, limiter and switching applications. Astable, monostable and bistable operations have been verified and binary and ring counters have been constructed with combinations of this silicon device and usual germanium diodes. Further circuit efforts are now going on. The development of new circuits appropriate for this low-impedance device will be necessary for wide applications.

Acknowledgements—The authors wish to express their particular thanks to G. M. HATOYAMA and W. SASAKI of Electrotechnical Laboratory for their help in measurements at liquid-helium temperature and also acknowledge their gratitude to K. IWAMA who encouraged them in the course of the study.

REFERENCES

1. L. ESAKI, *Phys. Rev.* **109,** 603 (1958).
2. L. ESAKI, *Properties of Heavily-Doped Germanium and Narrow p–n Junctions*. Presented at the International Conference on Solid State Physics in Brussels, (1958), Academic Press, London (1959).
3. J. J. EBERS, *Proc. Instn. Elect. Engrs.* (Supplement No. 15 or 17 to Part B). To be published.
4. T. YAJIMA and L. ESAKI, *J. Phys. Soc. Japan* **13,** 1281 (1958).
5. C. ZENER, *Proc. Roy. Soc.* **145,** 523 (1934).
6. K. B. MCAFEE, E. J. RYDER, W. SHOCKLEY and M. SPARKS, *Phys. Rev.* **83,** 650 (1951).
7. G. DRESSELHAUS, A. F. KIP and C. KITTEL, *Phys. Rev.* **98,** 368 (1955).
8. L. ESAKI, *IRE Wescon Convention Record, Part* 3, p. 176 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 22–33. Printed in Great Britain

PHYSICAL PRINCIPLES OF THE ESAKI DIODE AND SOME OF ITS PROPERTIES AS A CIRCUIT ELEMENT

R. A. PUCEL

Research Division, Raytheon Company, Waltham, Massachusetts

(*Received* 30 *June* 1959; *in revised form* 14 *August* 1959)

Abstract—A brief description is given of the physical mechanism underlying the operation of the recently invented, negative-resistance "tunnel" diode. This diode, which is capable of operation at microwave frequencies, is then analyzed from a circuit viewpoint. On the basis of this analysis, an equivalent circuit of the diode is obtained with the help of which are calculated such useful design criteria as the maximum frequency of oscillation and the maximum gain-bandwidth product, and the minimum noise figure, when the diode is used as an oscillator and amplifier, respectively. Experimental results are quoted which verify many of these quantities.

Résumé—On décrit rapidement le mécanisme physique qui permet le fonctionnement de la diode "tunnel" à résistance négative récemment inventée. Cet élément qui peut fonctionner même dans la gamme des micro-ondes est ensuite analysé du point de vue circuit. De cette analyse on déduit un circuit équivalent qui permet de calculer les critères importants tels que la fréquence maximum d'oscillations; le produit gain-largeur de bande et le facteur de bruit minimum, lorsqu'on utilise la diode comme oscillateur ou comme amplificateur respectivement. On décrit des résultats expérimentaux qui vérifient ces prévisions.

Zusammenfassung—Es wird eine kurze Beschreibung des physikalischen Mechanismus gegeben, der der kürzlich erfundenen ,,Tunnel''-Diode mit negativem Widerstand zu Grunde liegt. Diese Diode, die bei Ultrahochfrequenzen zu arbeiten imstande ist, wird vom Gesichtspunkt des Ersatzschaltbildes her analysiert. Es wird ein Ersatzschaltbild ermittelt, mit dessen Hilfe die maximale Oszillationsfrequenz, das maximale Produkt aus Verstärkung und Bandbreiten und das minimale Rauschverhältnis berechnet werden. Experimentelle Ergebnisse werden angeführt, die diese Berechnungen bestätigen.

INTRODUCTION

THE Esaki or "tunnel" diode[1,2] is a specially prepared semiconductor diode which exhibits a negative resistance over a limited (several tenths of a volt) voltage range when biased in the forward direction. The negative resistance is of the non-parametric type and arises from the so-called "tunneling mechanism", a strictly quantum-mechanical effect which will be described later.

Because of its possible (still unverified) low noise properties and high-frequency applicability as an amplifier or a low-level (mW) oscillator, and as a high-speed computer switch, and because of its inherent simplicity of fabrication, this device is being studied by numerous laboratories, including our own at the Research Division. Study of these devices is being advanced from both the theoretical and experimental viewpoint.

In a conventional p–n diode, conduction occurs predominantly by the diffusion of minority carriers across the p–n junction. In the forward-biased direction, these carriers consist of holes which diffuse from the p-side to the n-side, and electrons which travel in the opposite direction. Both contributions, because of the difference in the sign of the charge, constitute a conventional current of the same sign.

A new conduction process is introduced, however, when the two sides of the junction are doped so heavily with controlled impurities that the Fermi level E_F, instead of falling within the forbidden gap, as is usual, falls within the valence band on

the *p*-side and within the conduction band on the *n*-side (degenerate doping) as shown in Fig. 1 below. This causes an overlap Δ of the band edges. This new mechanism, a non-classical phenomenon, arises from the so-called "tunnel effect" or internal field emission first considered by ZENER[3], wherein electrons in a given energy state from the valence band can, without changing their energy state, "tunnel" across the forbidden gap into the empty or conduction band, and vice versa, under the action of a large electric field which causes a severe bending of the energy bands as shown in Fig. 1.

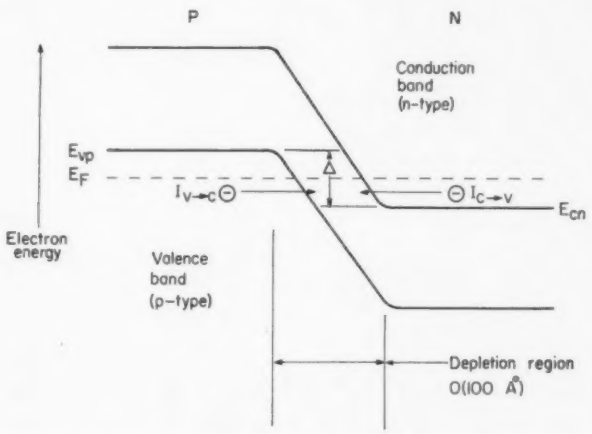


FIG. 1. Energy-band scheme at junction of an Esaki diode at thermal equilibrium.

It is not appropriate here to go into the quantum-physical details of the process. We shall be content to give only the general picture of the process.

As mentioned above, the only electrons which are involved are those falling in states included in the band overlap. Since charges of one sign are involved, the two components of current must be *subtracted* from each other. In particular, with no external bias applied, the net current must be zero; therefore, a mutual cancellation must take place, though each component by itself is non-zero (at temperatures above absolute zero). On the other hand, the net current must also vanish for a (forward) voltage which just destroys the overlap ($V = \Delta/q$), but in this case the individual components are also zero. For intermediate voltages, a net forward current flows because the rates of change of these two components with bias differ. The above facts necessarily imply a negative resistance over some voltage range. This will be illustrated below.

A FURTHER ELUCIDATION OF THE ORIGIN OF THE NEGATIVE-RESISTANCE REGION

Consider again the energy-band scheme for this diode, Fig. 1. (The energy-band scheme might not be entirely valid for this case because of the extremely steep potential gradients, of the order of 10^5–10^6 V/cm.) Following the development proposed by ESAKI, let the current caused by electrons tunneling from the valence to the conduction band in an energy state interval dE of the overlap region be labelled $-dI_{v\to c}$ (minus sign referring to the fact that this is a $-x$ directed conventional current). This current will be proportional to the number of valence electrons occupying the states in this energy interval, the number of electronic states in this same interval, in the conduction band, which are empty and are available for occupation, and the transition probability of tunneling from the valence to the conduction bands. Let $\rho_c(E)$ and $\rho_v(E)$ represent the energy-state densities in the conduction and valence bands, respectively, $f_c(E)$ and $f_v(E)$ the corresponding Fermi distribution functions denoting the probability that a given energy state is occupied, and $Z_{v\to c}$ and $Z_{c\to v}$ the

transition probabilities in the two directions. Then

$f_v(E)\rho_v(E)$ = density of valence-band electron states *occupied* in dE,

$(1-f_c)\rho_c(E)$ = density of conduction-band electron states *unoccupied* in dE.

Thus

$$I_{v\to c} = K\int_{E_{cn}}^{E_{vp}} Z_{v\to c}\rho_v(E)\rho_c(E)f_v(1-f_c)\,dE \quad (1)$$

where K is a constant. The interval of integration is from the conduction-band edge of the n-side, E_{cn}, to the valence-band edge of the p-side, E_{vp}, that is through the overlap. Likewise

$$I_{c\to v} = K\int_{E_{cn}}^{E_{vp}} Z_{c\to v}\rho_v(E)\rho_c(E)f_c(1-f_v)\,dE. \quad (2)$$

Assuming $Z_{c\to v} = Z_{v\to c} = Z$ as a convenient approximation, the net current is given by

$$I = I_{c\to v} - I_{v\to c} = K\int_{E_{cn}}^{E_{vp}} Z(f_c-f_v)\rho_c\rho_v\,dE \quad (3)$$

At thermal equilibrium, the probability that a given energy state E is occupied by electrons is the same whether this state falls within the valence band or the conduction band, that is $f_c(E) = f_v(E)$. Hence $I_{c\to v} = I_{v\to c}$ and $I = 0$, or both components mutually cancel. When a forward bias V is applied to the junction, the energy states, Fermi level, and band edges of the p-side shift downwards by an amount qV with respect to those of the n-side as shown in Fig. 2. (The Fermi levels should more correctly be called quasi-Fermi levels at nonthermal equilibrium.) Because the band overlap Δ changes upon application of a forward bias, each current component also changes. Clearly, when $V = \Delta/q$, the overlap is completely annihilated, tunneling ceases and both components reduce to zero.

To understand how each component varies between the two extremes $V = 0$, Δ/q, and in particular, how the negative resistance arises, it is convenient to consider the tunneling process at absolute zero since then the Fermi distributions are

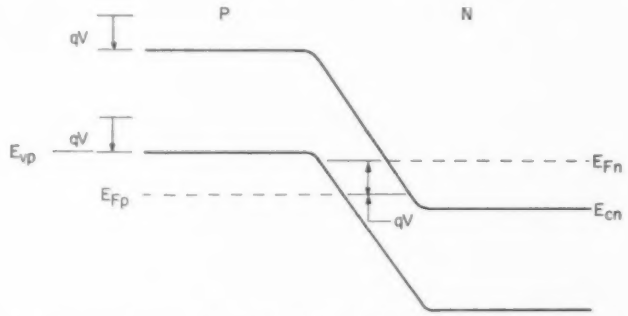


FIG. 2. Energy-band scheme of Esaki diode for an applied forward bias V.

simple step functions as shown in Fig. 3(a). On the p-side the distribution function $f_v(E)$ implies that all energy states below the quasi-Fermi level E_{Fp} are occupied and all above are empty. A corresponding statement holds for $f_c(E)$, E_{Fn} and the n-side.

In Fig. 3(b), the cross-hatched regions in the energy-band sketch denote the fully occupied states at zero bias. It is clear that since all states above E_F are empty in either band, both currents are zero. Upon application of a forward bias V, the band picture is changed to that of Fig. 3(c). It is evident that tunneling originating in the valence band is still non-existent and remains so for all values of forward bias, since all occupied valence states are adjacent to fully occupied conduction states. In other words $I_{v\to c}$ is identically zero for $V \geqslant 0$. On the other hand, tunneling does occur from the conduction band because some of the fully occupied conduction states are in line with empty valence states, thus $I_{c\to v} > 0$. However, since

$I_{c \to v}$ is again zero at $V = \Delta/q$, it must reach a maximum somewhere in the interval $0 < V < \Delta/q$. Since $I = I_{c \to v}$, I must do the same. For a reverse bias, the band picture changes to that of Fig. 3(d). Using reasoning analogous to that employed above, it is obvious that no tunneling from the conduction band can take place for any reverse bias

The situation at temperatures above absolute zero is not altered drastically. Because the Fermi functions are not discontinuous, but have a "transition region" of the order of several kT on either side of the respective quasi-Fermi level, as shown in Fig. 4(a), there are some states in each band which are partially occupied above this level and some

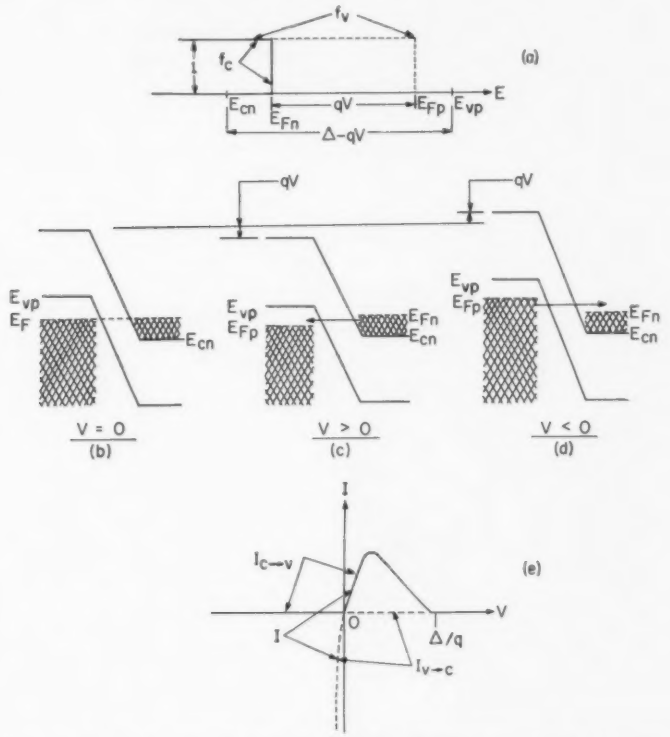


FIG. 3. Fermi functions, energy bands and tunnel currents at absolute zero for various bias voltages.

because all states are empty above E_{Fn} in the conduction band. Thus $I_{c \to v}$ is identically zero. Tunneling does occur, however, in the opposite direction and increases very rapidly with bias because of the enhanced overlap (origin of the well-known "Zener breakdown").

The above points are graphically illustrated in Fig. 3(e). Taking the difference of these two components, one obtains the net tunnel current, also shown in this figure. The origin of the resulting negative-resistance region clearly shows itself and is obviously due to $I_{c \to v}$.

which are partially empty below. Comparing with the previous case, a little thought will show that some tunneling can occur in either direction for $V = 0$. For a forward bias, the tunnel current originating in the conduction band has the general shape as at $T = 0$, i.e. a maximum at some value of V. The other tunnel current, though now not identically zero, decreases rapidly to zero as $V \to \Delta/q$. For a reverse bias, $I_{c \to v}$ is not identically zero as before, but rapidly decreases to this limit as the reverse bias is increased. The component $I_{v \to c}$ on the other hand varies in a similar

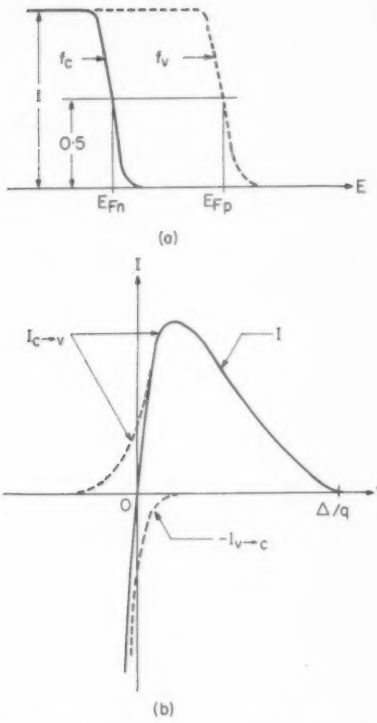


FIG. 4. The Fermi functions and tunnel currents for $T > 0$.

fashion as for $T = 0$. These statements are illustrated in Fig. 4(b). The net tunnel current is also shown. Compare Figs. 4(b) and 3(e).

Occurring concurrently with this tunneling is the ordinary diffusion process. The diffusion current can be quite small over the voltage range where tunneling occurs (of the order of several tenths of a volt in the forward-bias range). The composite V–I characteristic is illustrated in Fig. 5. If the temperature is so high that the diffusion process masks out the tunneling mechanism, no negative resistance will be exhibited by the diode.

PROPOSED EQUIVALENT CIRCUIT OF THE ESAKI DIODE

Based on the above considerations it is reasonable to assume that the simplest equivalent circuit of the diode in its active negative-resistance mode is as shown in Fig. 6. The negative resistance $-R$ or conductance $-G$ (R and G positive quantities) in parallel with the capacity C represents the junction itself, while the series resistance R_b represents the bulk and contact losses. Parasitic elements like lead inductances are not included. The parallel combination is justified from the standpoint that the tunnel current flows independently of the displacement current which is caused by the barrier

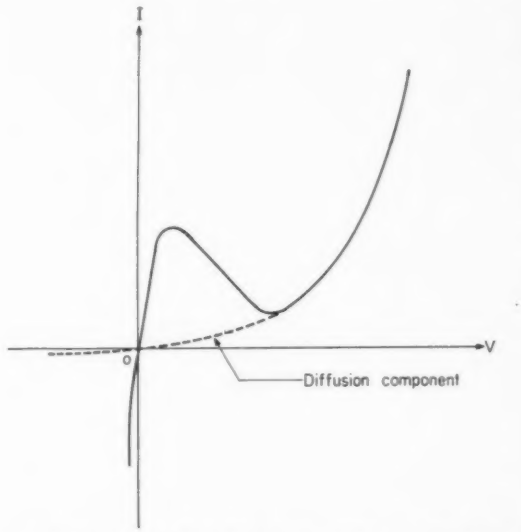


FIG. 5. Composite V–I characteristic of the Esaki diode including diffusion current.

widening and narrowing under the action of an applied alternating potential. [The capacity C also includes a frequency-dependent diffusion part of which is linearly dependent on the diffusion component of current. However, since this component in the negative-resistance range is very small, and since the barrier or depletion-layer capacity is so

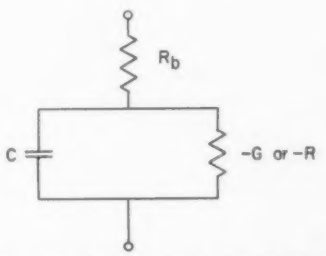


FIG. 6. Proposed equivalent circuit for Esaki diode exclusive of parasitic elements.

large because of the narrow barrier, C represents mainly the barrier (frequency-independent) capacity. Similarly, the negative conductance consists of the tunnel conductance and a positive conductance proportional to diffusion current in parallel with it. However, as for the diffusion capacity, the diffusion conductance is small and can be neglected in many cases.] The series resistance R_b is of an order of magnitude less than R, because of the highly-doped materials used.

If it were not for the series losses R_b, this diode would exhibit a negative resistance at its terminals at all frequencies and would *in principle* permit construction of oscillators and amplifiers at any frequency regardless of the size of C. Practically speaking, of course, the maximum tolerable value of C is limited by the smallest inductance (in lumped or distributed systems) which can be had. However, because R_b is *not* identically zero, there is a *theoretical* upper limit to the frequency at which the diode is useful, namely, that frequency at which the diode at its terminals no longer exhibits a negative resistance. It is because of R_b, that the shunting effect of C becomes theoretically important (and practically, of course).

This upper frequency limit, which we shall call the cut-off frequency f_c, is obtained by evaluating the frequency-dependent real part of the terminal impedance and setting it equal to zero to solve for f_c.

By computation

$$f_c = \frac{1}{2\pi RC}\sqrt{\left(\frac{R}{R_b}\right)-1} \tag{4a}$$

Since in practice $R \gg R_b$ for a well-made diode,

$$f_c = \frac{1}{2\pi C\sqrt{(R \,.\, R_b)}} = \frac{1}{2\pi C}\sqrt{\left(\frac{G}{R_b}\right)} \tag{4b}$$

At this frequency the net negative resistance of the junction is just cancelled by the series resistance R_b. This result clearly illustrates that for a diode to be a "good" diode, the series resistance and the shunt capacity must be as small as possible and the shunt conductance as large as possible.

If the shunt capacity C arises physically from precisely the same junction area over which tunneling occurs, then a reduction of the area of the junction has no effect on f_c, because both C and G are proportional—and R_b is inversely proportional to this area. However, since the tunnel current, and hence G, is a very strong function of the barrier thickness w, varying roughly as $w^{-1}e^{-kw}$ whereas C varies only as w^{-1}, then because of inevitable variations of the barrier thickness arising from deviations in impurity concentration over the junction cross-section, tunneling will occur predominantly over small areas where the thickness is a minimum. The capacity on the other hand, will not be affected very strongly by these deviations and will be determined by the entire physical area of the junction. It follows, therefore, that by reducing the physical area, that is doped point contact versus an alloy junction, one might be able to reduce the displacement current by a much larger factor than the tunnel current. Furthermore, the series resistance will also not be affected so strongly (though more so than G, it is felt) because it is determined primarily by the area over which the tunnel current emerges from the junction. A capacity of less than 10 $\mu\mu$F is easily attainable. Thus, a reduction in junction area should in practice yield an increase in cut-off frequency.*

Though the diode will present a negative conductance at all frequencies up to f_c, this conductance will be appreciable and essentially frequency-independent only for frequencies considerably below f_c. It is easily verified that the diode presents at its terminals an admittance consisting of an equivalent parallel capacity C_{eq}, and negative conductance of magnitude G_{eq},

$$Y = -G_{eq}+j\omega C_{eq} \tag{5}$$

where

$$G_{eq} = \left(\frac{1}{R-R_b}\right)\frac{1-(f/f_c)^2}{1+(f/f_c)^2[R_b/(R-R_b)]} \tag{6}$$

and

$$C_{eq} = C\frac{[R/(R-R_b)]^2}{1+(f/f_c)^2[R_b/(R-R_b)]} \tag{7}$$

For a good diode, $R \gg R_b$, thus for frequencies not greater than f_c,

$$G_{eg} \approx \frac{1}{R-R_b}\left[1-\left(\frac{f}{f_c}\right)^2\right] \approx G\left[1-\left(\frac{f}{f_c}\right)^2\right] \tag{8a}$$

and

$$C_{eq} \approx C \tag{8b}$$

* The above argument concerning area reduction was advanced to the author by G. RUPPRECHT of this Division.

The equivalent circuit of Fig. 6 can therefore be simplified to one with $-G_{eq}$ and C in parallel. This is very convenient for gain and noise figure calculations.

Though G_{eq} is frequency-dependent, if f is not too near f_c it may be considered a constant over any frequency interval small compared with f_c. In any derivations which follow this will be assumed. Furthermore, if $f \ll f_c$, G_{eq} may be replaced by its low-frequency value G. For example, when f is as high as $0{\cdot}5\,f_c$, $G_{eq} = 0{\cdot}75\,G$, when $f = 0{\cdot}25\,f_c$, $G_{eq} = 0{\cdot}94\,G$. [A method of measuring f_c indirectly might be as follows. Using the diode as an oscillator, one finds experimentally what external loading G_{eq} will just stop oscillations at various frequencies. Then plotting $(1-G_{eq}/G)$ versus frequency on log-log paper, where G is the threshold loading at a very low frequency, one can obtain f_c from an intercept of this curve.] For convenience we shall use henceforth, the symbol G for G_{eq} with the understanding that G might not coincide with the low-frequency value. The equivalent circuit to be used in the following discussion is shown in Fig. 7.

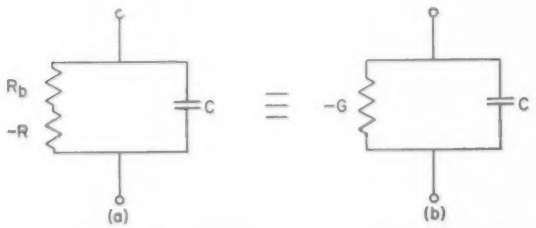


FIG. 7. Simplified equivalent circuit of the Esaki diode valid for frequencies below cut-off.

GAIN-BANDWIDTH PRODUCT OF THE ESAKI AMPLIFIER

The equivalent circuit of a negative-resistance amplifier using this diode is shown in Fig. 8. The current I_s and conductance G_s characterize the sinusoidal signal source. The conductance G_L represents the load, and L the tuning inductance.

Though there are many definitions of "power gain," since this is a negative-resistance amplifier, the most convenient one is the transducer gain G_t, (not to be confused with available gain, which is inappropriate here). This is defined as the ratio of the power actually delivered to the load P_0, divided by the maximum possible power which the given source can deliver to any load, namely, the available power $P_{av} = |I_s|^2/4G_s$. Thus

$$G_t = \frac{P_0}{P_{av}} = \frac{|I_0|^2/G_L}{|I_s|^2/4G_s} \tag{9}$$

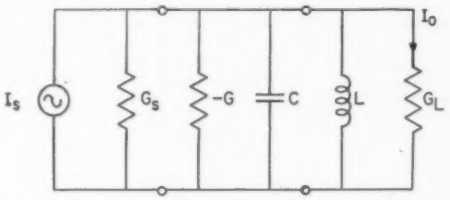


FIG. 8. Equivalent circuit of the Esaki amplifier.

From Fig. 8,

$$\frac{|I_0|^2}{|I_s|^2} = \left| \frac{G_L}{G_L+G_s-G+j\left(\omega C-\dfrac{1}{\omega L}\right)} \right|^2 = \frac{G_L^2}{(G_L+G_s-G)^2+\omega^2C^2\left(1-\dfrac{\omega_0^2}{\omega^2}\right)^2} \tag{10}$$

where $\omega_0 = \sqrt{[(LC)^{-1}]}$ is the resonant frequency. Therefore

$$G_t = \frac{4G_sG_L}{(G_L+G_s-G)^2+\omega^2C^2\left(1-\dfrac{\omega_0^2}{\omega^2}\right)^2} \tag{11}$$

At resonance, this gain is

$$G_0 = \frac{4G_sG_L}{(G_L+G_s-G)^2} \tag{12}$$

Clearly, as the external loading G_L+G_s approaches the negative conductance in magnitude, that is $G_L+G_s \to G$, the resonance gain becomes large without limit, and the amplifier degenerates into an oscillator.

The bandwidth of the amplifier is defined as the difference of the two frequencies adjacent to ω_0 at which G_t decreases to half its resonance value, that is when

$$\omega^2C^2\left(1-\frac{\omega_0^2}{\omega^2}\right)^2 = (G_L+G_s-G)^2 \tag{13}$$

This has four solutions, only two of which are positive, the other two being images on the negative–real frequency axis. The two positive values are given exactly (this is *not* a narrow-band approximation) by

$$\omega_{2,1} = \sqrt{\left[\left(\frac{G_s+G_L-G}{2C}\right)^2+\omega_0^2\right]}\pm\frac{G_s+G_L-G}{2C} \tag{14}$$

The bandwidth in c/s is

$$B=\frac{1}{2\pi}(\omega_2-\omega_1)=\frac{G_s+G_L-G}{2\pi C} \tag{15}$$

which shows clearly that the higher the resonance gain, the smaller the bandwidth and vice versa. A convenient measure for characterizing the amplifier is its gain-bandwidth product defined by

$$\sqrt{(G_0)}B=\frac{\sqrt{(G_sG_L)}}{\pi C} \tag{16}$$

which is independent of the gain and dependent only on C and the distribution of the loading between the source and load. The higher the gain-bandwidth product, the better the amplifier. This expression again points out the need for a small C.

The maximum gain-bandwidth product occurs at high gain $G_s+G_L \approx G$. The appropriate distribution of the loading at high gain for which this is achieved for a given diode occurs when the product

$$\sqrt{(G_sG_L)} \approx \sqrt{[G_L(G-G_L)]} \tag{17}$$

is a maximum. This happens for $G_L = G_s = G/2$, that is, for symmetric loading. Thus

$$[\sqrt{(G_0)}B]_{\max}=\frac{G}{2\pi C} \tag{18}$$

which depends only on the parameters of the diode itself. Clearly, the more pronounced is the tunneling effect, the higher the product. This product is a useful figure of merit for diode-design purposes. Notice that, theoretically, it is independent of the junction area. A sketch of the gain-bandwidth product at high gain as a function of source and load conductance is shown in Fig. 9.

Symmetric loading, however, is not necessarily useful for other purposes, such as minimization of the noise figure, as will be shown in the following section.

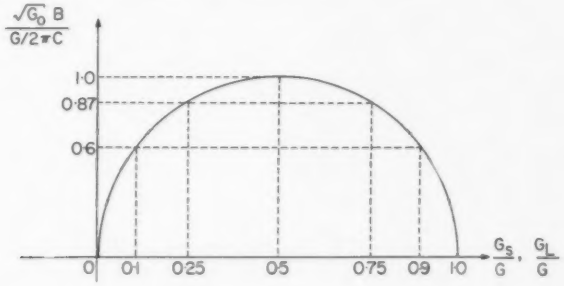


FIG. 9. Dependence of gain-bandwidth product on source and load conductance at high gain.

NOISE FIGURE

There are two sources of noise in the Esaki diode, a thermal noise generated by the series resistance and a noise produced by the random variations in the electron crossings at the junction. We believe the latter to be a shot noise by analogy with emission noise of vacuum tubes and ordinary diffusion diodes. Our experiments show that this shot-noise hypothesis is correct. This component of noise can be represented by a noise-current generator in parallel with the negative resistance $-R$ of the junction, Fig. 7(a). The thermal noise can be represented by a noise voltage in series with R_b. Quantitatively, these sources produce, in a small frequency range δf, a mean-square value of noise current and voltage, respectively, given by

$$\begin{aligned}\overline{i^2} &= 2\Gamma q\bar{I}\delta f\\ \overline{e_b^2} &= 4kTR_b\delta f\end{aligned} \tag{19}$$

where q is the electronic charge, k is Boltzmann's constant and T is the absolute temperature. The factor Γ is a constant less than or equal to unity which denotes whether the particle noise is a full shot noise ($\Gamma = 1$) or less. This depends on whether the fluctuations arising from individual passage of carriers across the junction are completely random or partially smoothed because of correlation such as occurs in vacuum tubes in the space-charge-limited regime.[4] Correlation between junction crossings may possibly result here from fluctuations in the barrier potential as each carrier crosses the junction, because of charges induced on the depletion-layer boundaries by the moving charge.*

The current $\bar{I}$ is not the actual net bias current but rather the sum of the *magnitudes* of currents flowing in both directions, since current fluctuations cannot be distinguished by the direction of current flow. Thus $\bar{I} = |I_{c \to v}| + |I_{v \to c}| + |I|_{\text{diffusion}}$. Since the diffusion current may be negligible under appropriate conditions (i.e. low enough temperature) and proper diode design, tunnel current will probably dominate $\bar{I}$. Furthermore, since in the negative-resistance region $I_{c \to v}$ is the dominant tunnel component, $\bar{I}$ should not differ too markedly from the bias current $I_{c \to v} - I_{v \to c} + I_{\text{diffusion}}$. The experimental study indicates this to be approximately true in the samples tested.

Based on the above hypotheses, the equivalent circuit to be used for noise-figure calculations at resonance is shown in Fig. 10. We shall compute noise contributed by the circuit exclusive of that generated by the load G_L or the stage which follows, such as a detector-in accordance with the commonly accepted definition of noise figure. The contribution of the load or second stage can be added to our result using the proposals advanced by H. Haus[5].

The noise powers delivered by each noise generator, including the thermal noise of the source, are proportional to the load conductance G_L. Since the noise figure F is just the sum of all noise powers delivered to the load divided by that contributed by the source, the quantity G_L cancels out. Thus, one may compute F using any value of G_L, in particular $G_L = 0$. We shall do so here for convenience. Therefore

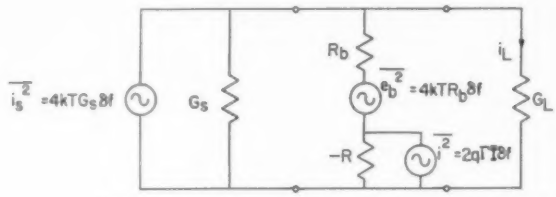


FIG. 10. Circuit used for computation of first-stage noise figure.

$$F = \frac{\text{total noise power delivered to } G_L \text{ by all sources}}{\text{noise power delivered to } G_L \text{ by source } i_s} = \overline{i_L^2}/\overline{i_s^2} \tag{20}$$

Referring to Fig. 10, with $G_L = 0$, the contribution to $\overline{i_L^2}$ by the various sources is easily shown to be

$$\overline{i_L^2} = \overline{i_s^2} + \frac{\overline{e_b^2}}{(R-R_b)^2} + \overline{i^2}\left(\frac{R}{R-R_b}\right)^2 \tag{21}$$

Then

$$F = \frac{\overline{i_L^2}}{\overline{i_s^2}} = 1 + \frac{\overline{e_b^2}}{\overline{i_s^2}(R-R_b)^2} + \frac{\overline{i^2}}{\overline{i_s^2}} \cdot \left(\frac{R}{R-R_b}\right)^2 \tag{22}$$

Assuming room temperature for all noise sources, we obtain

$$F = 1 + \frac{R_b}{G_s(R-R_b)^2} + \left(\frac{R}{R-R_b}\right)^2 \frac{\bar{I}\Gamma}{2G_s(kT/q)} \tag{23}$$

* Suggested by H. Statz of this Division.

For a well-made diode $R—R_b \approx R = G^{-1}$, thus

$$F = 1+\frac{R_b}{R}\left(\frac{G}{G_s}\right)+\frac{\bar{I}\Gamma}{2G_s(kT/q)}, \quad (R = G^{-1}) \tag{24}$$

(If the load noise is included, and if it consists simply of a thermal nature, an additional term G_L/G_s should be included.) The second term should be negligible for any well-designed diode.

On the other hand, for symmetric loading (which gives the maximum gain-bandwidth product),

$$F_s = 1+2\frac{R_b}{R}+\frac{\bar{I}\Gamma}{GkT/q} \tag{26}$$
$$= 1+2\frac{R_b}{R}+40\Gamma\bar{I}R \text{ at room temperature.}$$

Comparing (25) and (26), it is clear that the symmetric excess-noise figure $F_s—1$ is just double the minimum possible value.

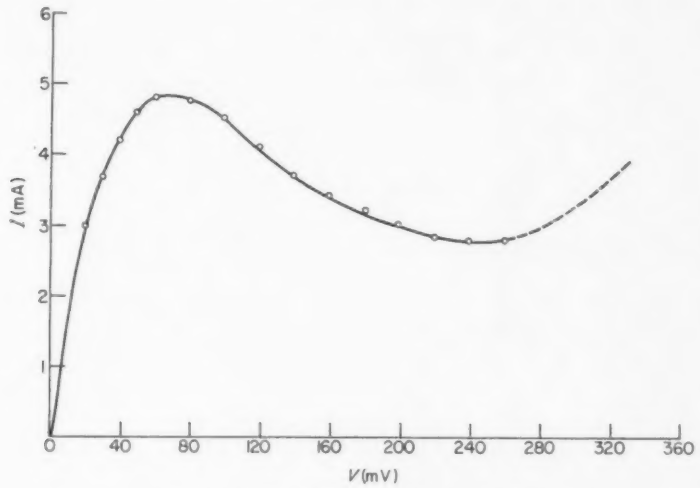


FIG. 11. I–V characteristic of Esaki diode in forward-biased region.

For a given diode, it is clear that the higher G_s is, the lower F is. In other words, F is a minimum for the non-symmetric loading case, for which the dominant loading is produced by the source. This has been found elsewhere for parametric amplifiers also. If the amplifier is to operate between fixed loads, say $G_s = G_L$, the asymmetry can be achieved by using appropriate transformers at either end.

Since for high gain, the maximum possible value of G_s is just G (as $G_s+G_L \approx G$ for high gain), then the minimum value of F is

$$F_{\min} = 1+\frac{R_b}{R}+\frac{\bar{I}\Gamma}{2G(kT/q)} \tag{25}$$
$$= 1+\frac{R_b}{R}+20\Gamma\bar{I}R \text{ at room temperature.}$$

EXPERIMENTAL RESULTS*

A typical I–V characteristic obtained with an X–Y recorder of one of the samples (silicon) tested is shown in Fig. 11. The maximum slope found graphically from this curve, corresponds to a negative resistance slightly less than 50 Ω.

The diode capacitance C was measured by noting the shift in oscillation frequency when a known capacitor was placed in parallel with the diode. The value so obtained for the above diode (No. 84) was $C = 500\ \mu\mu F$. This parameter could also have been measured with an impedance bridge provided the diode was loaded sufficiently to render it

* The entire experimental program was conducted by S. COHEN of this Division.

passive. (These parameters should not be considered typical of what can be realized.)

Using ESAKI's estimate for the bulk resistance of this particular diode, $R_b \approx 10^{-2}\ \Omega$, we calculate a cut-off frequency of 450 Mc/s. However, because of the unusually large capacitance, this diode was restricted to a much lower limit of frequency operation by the self-resonance with the package lead inductance.

The circuit used to measure gain-bandwidth and noise at 30 Mc/s is shown in Fig. 12. The 0·5-Ω resistor is used to provide a low d.c. source impedance for the diode to prevent relaxation oscillations. Gain was varied by adjusting the 100-Ω potentiometer in series with the signal generator.

Table 1. R_s (*for maximum gain*) $= 62\ \Omega$; R_L (*corrected for tank loading*) $= 890\ \Omega$; $C = 500\ \mu\mu$F; $V = 0{\cdot}170$ V; I (*bias current*) $= 3{\cdot}3$ mA

Diode No. 84 Gain (db)	B (Mc/s)	$\sqrt{(G_0)}B$ (Mc/s)	
		(Measured)	(Calculated)
28	0·13	3·2	2·7
24	0·20	3·2	2·7
21	0·27	3·0	2·7
18	0·40	3·2	2·7
16	0·49	3·1	2·7
14	0·61	3·1	2·7

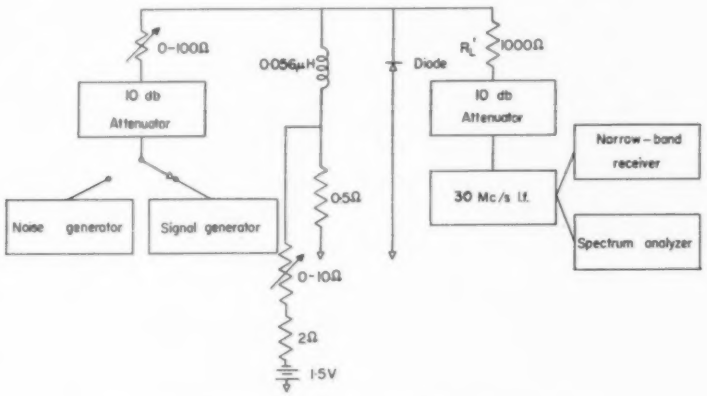


FIG. 12 Circuit for noise and gain measurement.

Bandwidth was measured for various values of gain. Gain was varied over a large range by slightly adjusting the source impedance potentiometer. The load was considered to consist of the kΩ resistance and all losses contributed by the tuning inductance and the d.c. source resistance in series with it. The measured and calculated gain-bandwidth products are listed in Table 1 below.

The maximum gain-bandwidth product occurs for symmetric loading of the amplifier ($R_s = R_L$). This was verified by operating the same diode with more symmetric loading, see Table 2.

The agreement obtained between the measured and theoretical values for the gain-bandwidth product justifies the use of the simple equivalent circuit given in Fig. 7(b).

A problem associated with the measurement of noise in a negative-resistance amplifier is the sensitivity of gain to variations in the noise source impedance [see equation (12)]. Fortunately, the temperature-limited, noise-diode source provides

Table 2. $R_s = 63\ \Omega$; $R_L = 140\ \Omega$ (*corrected for tank loading*); $V = 0{\cdot}150$ V; $I = 3{\cdot}6$ mA; $C = 545\ \mu\mu$F (*varies somewhat with operating point*)

Gain (db)	B (Mc/s)	$\sqrt{(G_0)}B$ (Mc/s)	
		(Measured)	(Calculated)
27·5	0·24	5·8	6·2
22·3	0·46	6·0	6·2
18·0	0·75	6·0	6·2

sufficient power at 30 Mc/s to permit the insertion of a 10 db isolation pad between it and the Esaki amplifier. Thus, this troublesome problem was minimized.

Another noise-measurement problem with negative-resistance amplifiers is associated with the fact that the noise of the load or second stage used for detection is also amplified by the first stage being tested. This contribution of the second stage must either be known or eliminated from the noise measurement in conformity with the definition of noise figure. We have done the latter here. The measured noise figure of the A.I.L. 30 Mc/s preamplifier used in the second stage was 2 db at a 50-Ω impedance level. However, the contribution of this amplifier to the overall first-stage amplifier noise was made negligible when the former was connected in series with a large adjustable load resistance (R_L' in Fig. 12).

Table 3 presents a summary of the measured and calculated noise figures for various source and load impedances, R_s and R_L respectively (the load impedance R_L includes the loading of R_L' and the losses of the tuning inductance and d.c. source), and diode bias currents I. The calculated values were made on the basis of full shot noise ($\Gamma = 1$) and with the assumption that the diode current I is nearly equal to the total particle current $\bar{I}$. All measurements were made at a gain level of 25 db.

Table 3

I (mA)	R_s (Ω)	R_L (Ω)	Noise figure (db)*	
			(Measured)	(Calculated)
2·9	200	890	11·2	11·0
3·3	96	890	8·0	8·7
3·6	72	890	7·6	7·9
3·9	57	890	7·3	7·4
4·1	53	890	7·0	7·3
3·9	150	140	11·1	11·0
4·5	132	140	11·7	11·1

* Excluding load-noise contribution.

The agreement between measured and calculated values is sufficiently good to justify the full shot-noise assumption. Observe that the noise figure becomes a minimum as the loading approaches the asymmetrical case $R_L \gg R_s$ in accordance with equations (24) to (26).

The above noise figures are not particularly commendable as a basis for a low-noise amplifier. However, they should not be considered as the best that one can realize with proper design and choice of materials.

To get some idea of the direction to be pursued in this matter, one may proceed as follows: from equation (25) it is seen that noise figure will be a minimum the lower is the product $\bar{I}R = \bar{I}/G$. If we approximate $\bar{I}$ by the diode current I, then this product at any operating point (I_0, V_0) is simply the difference $V_i - V_0$, where V_i is the voltage value at which the tangent to the V–I characteristic at the operating point intersects the V-axis. It is plausible to assume that this voltage difference will decrease as the voltage at which the current minimum occurs is decreased. This implies that a small band overlap is desirable.

On the other hand, this condition is inconsistent with the requirement of as large a linear dynamic operating region as possible. Obviously, a compromise may be necessary between these two conditions and with others, such as a prescribed impedance level.

Acknowledgements—The author wishes to thank L. ESAKI who so kindly supplied him with diodes and W. C. DUNLAP who obtained them. Special gratitude is expressed to my colleague, Mr. S. COHEN, who was responsible for obtaining the measurements. Appreciation is made to my numerous colleagues for many helpful suggestions and to Miss J. LILLIBRIDGE for typing this manuscript.

Note added in revision: The equivalent circuit and the formulae based on it have been proposed by others at the recent (1959) Solid State Device Research Conference held at Cornell University and found to be applicable. See also: H. S. SOMMERS, JR., *Proc.I.R.E.* **47**, 1201 (1959).

REFERENCES

1. L. ESAKI, *Phys. Rev.* **109**, 603 (1958).
2. L. ESAKI and Y. MIYAHARA, *Solid-State Electron.* **1**, 13 (1960).
3. C. ZENER, *Proc. Roy. Soc.* **145**, 523 (1934).
4. K. SPANGENBERG, *Vacuum Tubes* p. 308. McGraw-Hill, New York (1948).
5. H. HAUS. Private communication.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 34–38. Printed in Great Britain

ZUR EXPERIMENTELLEN BESTIMMUNG DER THERMOELEKTRISCHEN EFFEKTIVITÄT VON HALBLEITERN

ULRICH BIRKHOLZ*

Laboratorium der Zentralwerksverwaltung der Siemens-Schuckertwerke AG, Nürnberg

(*Received* 30 *May* 1959)

Zusammenfassung—Die technische Ausnutzung der thermoelektrischen Effekte ist in den letzten Jahren durch die Entwicklung von speziellen Halbleitern[1,2] möglich geworden. Der Wirkungsgrad von thermoelektrischen Generatoren, die Leistungsziffer von elektrothermischen Kühlaggregaten und der Heizkoeffizient bei der reversiblen Heizung werden durch die gleiche Materialkonstante, die Effektivität z, bestimmt. Diese Grösse errechnet sich aus absoluter Thermokraft α, elektrischer Leitfähigkeit σ und Wärmeleitfähigkeit κ nach der Beziehung

$$z = \frac{\alpha^2 \sigma}{\kappa}$$

Im folgenden wird ein Verfahren beschrieben, mit dem sich α, σ und κ im Zimmertemperaturbereich mit Hilfe einer einzigen Apparatur ermitteln lassen, wobei auftretende Messfehler durch Wärmestrahlung sowie Inhomogenität der Probestäbe diskutiert werden.

Abstract—The technical use of the thermoelectric effects has been practicable for some years, since the development of special semiconductors. The efficiency of thermoelectric generators, as well as the coefficient performance of Peltier refrigerators and the coefficient for reversible heating, are determined by the same material constant, the figure of merit z. This magnitude is obtained by absolute thermoelectric power α, electric conductivity σ and thermal conductivity κ according the equation

$$z = \frac{\alpha^2 \sigma}{\kappa}$$

In the following paper a method is described by means of which α, σ and κ are ascertainable in the range of room temperature by use of a single piece of apparatus; errors of measuring arising by heat radiation, as well as inhomogeneity of the samples, are discussed.

Résumé—L'utilisation technique de l'effet thermoélectrique est devenue possible ces dernières années grâce à la découverte de nouveaux semiconducteurs. L'efficacité des générateurs thermoélectriques, le rendement des groupes de refroidissement électrothermique et le coefficient de chauffage réversible sont caractérisés par une seule constante: le facteur de mérite z. Cette valeur se calcule à partir du pouvoir thermoélectrique absolu α, de la conductibilité électrique σ et de la conductibilité thermique κ d'après l'équation:

$$z = \frac{\alpha^2 \sigma}{\kappa}$$

On décrit ensuite une expérience dans laquelle on peut déterminer à l'aide d'un appareil unique α, σ et κ à la température ordinaire. On discute les erreurs de mesure provenant du rayonnement de chaleur ainsi que de l'inhomogénéité des échantillons.

* Siemens-Schuckertwerke AG, Abt. ZW/LS Nürnberg, Katzwanger Str. 150.

1. BESCHREIBUNG DER VERSUCHSANORDNUNG

Der Probestab wird zwischen zwei wassergekühlte Kupferelektroden geklemmt, deren Temperaturen mit Hilfe von zwei Thermostaten konstant gehalten werden können (Abb. 1).

Die Elektroden sind elektrisch isoliert durch die Grundplatte geführt, so dass sie gleichzeitig als Stromzuführung dienen könen. Zur Gewährleistung eines hinreichend guten Kontaktes befinden sich zwischen Elektroden und Probe Indiumfolien. Um Wärmeverluste durch Konvektion und Wärmeleitung der Luft zu vermeiden, ist der Anschluss an eine Hochvakuumpumpe vorgesehen. Verwendet werden im allgemeinen zylindrische Stäbe mit 5–10 mm Durchmesser und 40–60 mm Länge; jedoch sind auch Proben mit rechteckigem Querschnitt geeignet. Zur Bestimmung der Temperaturverteilung dienen drei dünne Manganin-Konstantan-Thermoelemente (Drahtdurchmesser 0,1 mm), deren Thermospannung an einem fünfstufigen Diesselhorstkompensator gemessen wird. Da sich die meisten Halbleiter nicht direkt löten lassen, werden zur Anbringung der Thermoelemente zunächst 0,1 mm starke Rillen in den probestab gedreht, in die dünne verzinnte Schlingen aus Kupferdraht gelegt werden. An diese Drahtschlingen können die Thermoelemente dann direkt angelötet werden.

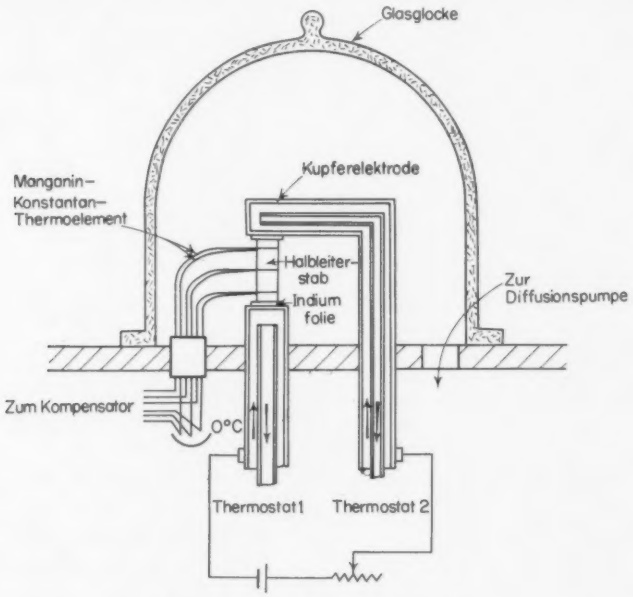


Abb. 1. Schema der Mesapparatur.

2. DIFFERENTIELLE THERMOKRAFT

Zur Bestimmung der differentiellen Thermokraft wird an dem Probestab eine Temperaturdifferenz von etwa 15 Grad eingestellt. Die Temperatur wird mit den drei Thermoelementen gemessen, deren Manganindrähte auch zur Bestimmung der Thermospannung des Halbleiters verwendet werden. Da die absolute Thermospannung von Manganin sehr gering ist (ca. 2 μV/Grad), erhält man auf diese Weise die absolute Thermokraft des Halbleiters. Bei inhomogenen Halbleiterpräparaten kann das Ergebnis der Messung von der Richtung des Wärmestromes abhängen. Dies ist auf ein nichtlineares Temperaturgefälle im Stäbchen infolge Wärmeabstrahlung zurückzuführen.

Für die an einem Stab der Länge l auftretende Thermospannung E gilt die allgemeine Beziehung

$$E(l) = \int_0^l \alpha(x) \frac{dT}{dx}\, dx$$

Wir setzen eine lineare Änderung der Thermokraft längs der Probe voraus, wie es in erster Näherung bei Präparaten der Fall ist, die durch gerichtetes Erstarren (normal freezing) hergestellt worden sind. Dem linearem Temperaturabfall am Probenstab sei eine parabolische Temperaturverteilung (Näherung) überlagert, die der Wärmeabstrahlung Rechnung trägt (Abb. 2).

Der Temperaturgradient lässt sich dann in folgender Form darstellen:

$$\frac{dT}{dx} = \frac{\Delta T}{l} - \frac{4\theta_{\max}}{l^2}(l-2x)$$

$\theta_{\max}$ ist die maximale Abweichung von der Linearität des Temperaturgefälles infolge Wärmeabstrahlung.

Ist $\Delta\alpha$ die Zunahme der Thermokraft längs des Probestabes, so gilt

$$\alpha(x) = \alpha_0 + \Delta\alpha\,\frac{x}{l} \quad \text{(1. Richtung des Wärmestromes)}$$

$$\alpha(x) = \alpha_0 + \Delta\alpha\,\frac{1-x}{l} \quad \text{(2. Richtung des Wärmestromes)}$$

Die Integration ergibt

$$E_{1\cdot2}(l) = \bar{\alpha}\Delta T \pm \tfrac{2}{3}\theta_{\max}\Delta\alpha$$

Die mittlere Thermokraft

$$\bar{\alpha} = \frac{1}{l}\int_0^l \alpha(x)\,dx$$

bestimmt sich aus dem Mittelwert beider Thermospannungen:

$$\bar{\alpha} = \frac{E_1 + E_2}{2\Delta T}$$

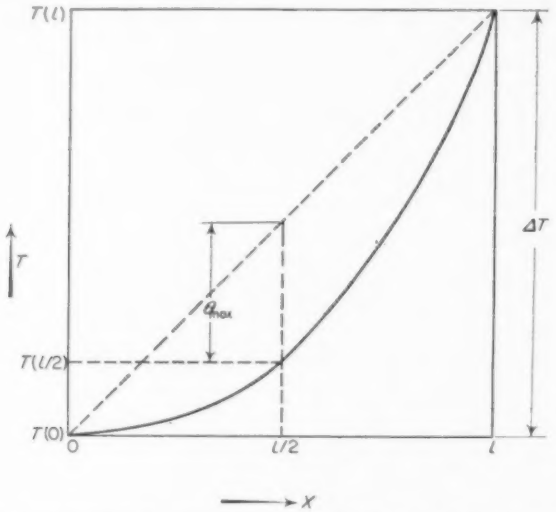


Abb. 2. Nichtlineares Temperaturgefälle bei der Thermokraftmessung.

3. ELEKTRISCHE LEITFÄHIGKEIT UND WÄRMELEITFÄHIGKEIT

Nach der von Diesselhorst[3] angegebenen Absolutmethode wurde das Verhältnis von Wärmeleitfähigkeit zu elektrischer Leitfähigkeit κ/σ und gleichzeitig die elektrische Leitfähigkeit gemessen, so dass die Wärmeleitfähigkeit berechnet werden konnte. Das Prinzip der Methode beruht auf der Messung der Temperatur und Potentialverteilung an einem elektrisch erwärmten Leiter. Der Probestab wird von einem elektrischen Strom j_e durchflossen. Wird die Temperatur der Elektroden so gewählt, dass die beiden äusseren Thermoelemente die gleiche Temperatur T_1 anzeigen, dann stellt sich am mittleren Thermoelement eine höhere Temperatur T ein. Temperatur und Potentialverlauf sind aus Abb. 3 ersichtlich. Die Spannungen U und V erhält man aus den gemessenen Werten durch Subtraktion der entsprechenden

Thermospannungen, die aus den zuvor gemessenen Thermokräften errechnet werden.

Nach Diesselhorst ergibt sich unter Vernachlässigung der thermoelektrischen Effekte und der Wärmeverluste durch Abstrahlung

$$\frac{\kappa}{\sigma T} = \frac{V(U-V)}{T^2-T_1^2}$$

Infolge der Wärmeverluste durch Abstrahlung wird die Differenz $T-T_1$ zu klein gemessen.

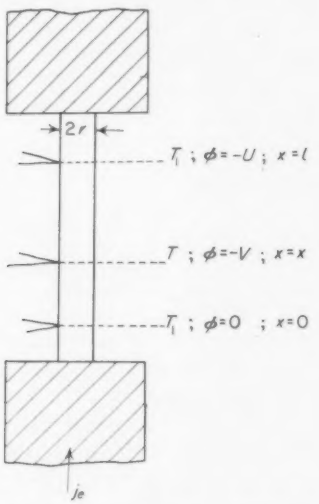


Abb. 3. Zuordnung von Temperatur und Potential an einem elektrisch erwärmten Leiter.

Ausserdem wird bei Proben, deren Thermokraft längs des Stabes nicht konstant ist, ein weiterer von der Stromrichtung abhängiger Fehler die Messung beeinflussen, da ein elektrischer Strom beim Durchfliessen eines Körpers mit variabler Thermokraft Peltierwärme entwickelt bzw. absorbiert.

Zur Verallgemeinerung obiger Formel dient der folgende Ansatz:

Wärmestrom $j_w = -\kappa\pi r^2 \frac{\partial T}{\partial x}$

elektrischer Strom

$$j_e = -\sigma\pi r^2 \frac{\partial \phi}{\partial x};\ \operatorname{div} j_e = 0$$

$$\operatorname{div} j_w = \frac{\partial j_w}{dx} = -j_e \frac{\partial \phi}{\partial x} + s(x) + p(x)$$

$s(x)$ Strahlungswärme je Längeneinheit
$p(x)$ Peltierwärme je Längeneinheit

Die Integration der Differentialgleichung ergibt:

$$\kappa\pi r^2 \frac{\partial T}{\partial x} = -j_e\phi + \int (s+p)\,dx + C$$

$$\frac{\kappa}{\sigma}\frac{\partial T}{\partial \phi} = -\phi + \frac{1}{j_e}\int_0^{\phi} (s+p)\frac{dx}{d\phi}\,d\phi + C_1$$

$$= -\phi + C_1 + f(\phi)$$

$$f(\phi) = \frac{1}{j_e}\int_0^{\phi} (s+p)\frac{dx}{d\phi}\,d\phi.$$

Das Wiedemann-Franz-Lorenzsche Gesetz $\kappa/\sigma T = \text{const}$ gilt nur für Metalle. Hier sei vorausgesetzt $\kappa/\sigma = \text{const}$, was für kleine Temperaturdifferenzen von 5 bis 10 Grad eine hinreichende Näherung ist.

$$\frac{\kappa}{\sigma}(T-T_1) = -\tfrac{1}{2}V^2 - C_1 V + \int_0^{-v} f(\phi)\,d\phi.$$

Mit der Randbedingung $T = T_1;\ V = U$ erhält man

$$\frac{\kappa}{\sigma} = \frac{V(U-V)+A}{2(T-T_1)}$$

$$A = -\frac{2V}{U}\int_0^{-u} f(\phi)\,d\phi + 2\int_0^{-v} f(\phi)\,d\phi.$$

Das Korrekturglied besteht aus einem Strahlungsanteil und einem Anteil, der die erzeugte Peltierwärme je Längeneinheit berücksichtigt:

$$A = A_s + A_p$$

Berechnung der Strahlungskorrektur A_s

Die Strahlungswärme je Längeneinheit beträgt bei der Umgebungstemperatur T_0 und dem relativen Emissionsvermögen ϵ

$$s(x) = -\epsilon\gamma 2\pi r(T^4 - T_0^4)$$

γ absolutes Emissionsvermögen des schwarzen Körpers
ϵ relatives Emissionsvermögen der Probe
T_0 Umgebungstemperatur

Es kann ein parabolischer Temperaturverlauf längs der Probe angenommen werden. Mit $T = T_0+\theta$ gilt

$$\theta = \frac{4\theta m}{l^2}(l-x)x+\theta_0;\ \theta_0 = T_1-T_0,$$

$\theta_m = T_m-T_1$; T_m = maximale Temperatur der Probe (Stabmitte). In guter Näherung wird $s(x)$ linearisiert:

$$s(x) \approx -\epsilon\gamma 2\pi r 4 T_0{}^3\theta$$

Man erhält

$$A_s = -\epsilon\gamma 2\pi r l \frac{U}{j_e} T_0{}^3[\tfrac{8}{6}(T-T_1)+(T_1-T_0)]$$

Durch einen Vorversuch ohne elektrischen Strom kann das Verhältnis von relativem Emissionsvermögen zur Wärmeleitfähigkeit ϵ/κ bestimmt werden. Die Probenenden werden auf eine Temperatur $T_1 > T_0$ gebracht, wobei sich in der Stabmitte infolge der Wärmeabstrahlung eine tiefere Temperatur T^* einstellt.

Die Lösung der entsprechenden Differentialgleichung ergibt mit

$$\Phi = \frac{a^2}{T_1{}^* 3}$$

$$\frac{\epsilon}{\kappa} = \frac{r}{2\gamma l^2}\cdot\Phi$$

Die Grösse a bestimmt sich aus den Messwerten nach der Gleichung

$$\cos \text{hyp}\, a = \frac{1}{1-4T_1{}^{*3}(T_1{}^*-T^*)/(T_1{}^{*4}-T_0{}^4)}$$

$$A_s = -\frac{\kappa}{\sigma}\Phi T_0{}^3[\tfrac{8}{6}(T-T_1)+(T_1-T_0)]$$

Berechnung der Peltierkorrektur A_p

Die pro Längeneinheit erzeugte bzw. absorbierte Peltierwärme beträgt

$$p(x) = -j_e T\frac{d\alpha}{dx}$$

Es sei wieder—wie bei der Diskussion der Thermokraftmessung—eine lineare Änderung von α vorausgesetzt:

$$\frac{d\alpha}{dx} = \text{const} = \frac{\Delta\alpha}{l}$$

Entsprechend der Rechnung für A_s ergibt sich

$$A_p = -\frac{\Delta\alpha U}{4}[\tfrac{8}{6}(T-T_1)+T_1]$$

$\Delta\alpha$ ist positiv zu nehmen, wenn der konventionelle Strom für eine p-leitende Probe in Richtung steigender Thermokraft fliesst. Die Zusammenfassung

$$\frac{\kappa}{\sigma} = \frac{V(U-V)+A_s+A_p}{2(T-T_1)}$$

liefert die Endformel

$$\frac{\kappa}{\sigma} = \frac{V(U-V)-\Delta\alpha U T'/4}{2(T-T_1)}\times$$

$$\times\frac{1}{1+\phi T_0{}^3(T'-T_0)/2(T-T_1)}$$

mit $T' = \frac{5}{6}(T-T_1)+T_1$.

Bei der praktischen Ausführung der Effektivitätsbestimmung wird zunächst die Thermokraft α durch Mittelung aus den Messungen für beide Richtungen des Wärmestromes erhalten.

Im Vorversuch (aufgeheizte Probenenden, kein Belastungsstrom) wird die Grösse ϕ bestimmt, die den Einfluss der Wärmeabstrahlung berücksichtigt.

Messung der Temperatur- und Potentialverteilung für beide Richtungen des elektrischen Stromes gestattet durch Eliminierung von $\Delta\alpha$ die Bestimmung von κ/σ, so dass z errechnet werden kann.

Aus der Spannung U, dem Strom j_e und der Geometrie kann die elektrische Leitfähigkeit berechnet werden. Damit ist auch die Wärmeleitfähigkeit κ bekannt.

Für die vollständige Bestimmung der thermoelektrischen Effektivität werden einige Stunden benötigt, die Unsicherheit der Ergebnisse ist kleiner als 3 Prozent.

LITERATUR

1. A. F. Joffe, *Halbleiterthermoelemente*. Moskau-Leningrad (1956).
2. U. Birkholz, *Z. Naturf.* **13a**, 780 (1958).
3. H. Diesselhorst, *Z. InstrumKde* **22**, 115 (1902).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 39–45. Printed in Great Britain

DIE THERMODYNAMISCHEN DATEN UND DAMPFDRUCKE* DER WICHTIGSTEN III–V-VERBINDUNGEN

THEODOR RENNER

Forschungslaboratorium der Siemens-Schuckertwerke AG, Erlangen

(*Received* 30 *May* 1959)

Zusammenfassung—Die bisher bekannt gewordenen thermodynamischen Daten wie Bildungsenthalpie, freie Bildungsenergie und Entropie der wichtigsten III–V-Verbingdungen werden in Tabellen zusammengefasst. Dampfdruckkurven für InAs, GaAs und InP werden mitgeteilt.

Abstract—Tables of the thermodynamic properties, such as enthalpy of formation, free energy of formation, and entropy, for the most important III–V-compounds have been prepared from data collected from various sources. Vapour-pressure data of InAs, GaAs and InP are presented.

Résumé—Les données thermodynamiques connues jusqu'à présent comme l'entropie de formation, énergie libre de formation et entropie des composés III–V les plus importants sont réunies en tableaux. On donne les courbes de pression de vapeur pour InAs, GaAs et InP.

1. EINLEITUNG

Die Kenntnis der thermodynamischen Daten der III–V-Verbindungen, wie Bildungsenthalpie, Entropie, freie Bildungsenergie und Dampfdruck ist sowohl für die Technologie als auch für die Deutung mancher Eigenschaften der III–V-Verbindungen von grossem Interesse. Im folgenden werden die bisher bekannt gewordenen Daten der wichtigsten Verbindungen in Tabellen zusammengefasst, sowie Kurven für den Dampfdruck von InAs, GaAs und InP gegeben.

2. BILDUNGSENTHALPIE DER III–V-VERBINDUNGEN

2.1 *Bildungsenthalpie bei Temperaturen* $> 298°K$

Es liegen Werte für InSb, InAs, GaAs, GaSb

Tabelle 1. *Bildungsenthalpie der* III–V-*Verbindungen bei* $T > 298°K$

Nr.	$A^{III}B^{V}$	Temp. (°K)	$\Delta W^{T}_{Bild.}$ (kcal/mol)	
1	InSb	723	−8,64	In fest, Sb fest (Ref. 7)
2	GaSb	900	−41,0	Ga flüssig, Sb_2 Gas (Ref. 4)
3		900	−25,5	Ga flüssig, Sb_4 Gas (Ref. 4)
4	InAs	1000	−43,1	In flüssig, As_2 Gas (Ref. 4)
5		1000	−26,0	In flüssig, As_4 Gas (Ref. 4)
6	InAs	~ 1000	$-39 \pm 2{,}5$	In flüssig, As_2 Gas (Ref. 14)
7		~ 1000	$-22 \pm 2{,}5$	In flüssig, As_4 Gas (Ref. 14)
8	GaAs	1080	−39,4	Ga flüssig, As_2 Gas (Ref. 5)
9		1080	−21,8	nach Ref. 5 berechnet; Ga flüssig, As_4 Gas
10	InP	1000	−39,4	In flüssig, P_2 Gas (Ref. 5)
11		1000	−25,4	nach Ref. 5 berechnet; In flüssig, P_4 Gas
12		1273	−26,0	In flüssig; P_4 Gas (Ref. 10)

* Unter „Dampfdruck" ist hier und im folgenden der Mindestdampfdruck zu verstehen, der erforderlich ist, um eine Dissoziation zu verhindern.

und InP vor. Sie beziehen sich teils auf feste, flüssige oder gasförmige Ausgangselemente der dritten bzw. fünften Gruppe des Periodensystems. In Tabelle 1 sind die Daten aufgeführt mit Quelle und Bezugsreaktion. Die Bildungsenthalpie ist—ideales Gasverhalten vorausgesetzt—druck*un*abhängig.

einem Zinnkalorimeter sowie von SCHNEIDER und KLOTZ fur InSb.[15] Für GaAs und InP lässt er sich näherungsweise aus den massenspektrometrisch[4,5,14] bei höheren Temperaturen bestimmten Werten berechnen, wenn man die molaren Wärmen der III–V-Verbindungen als konstant annimmt. In Tabelle 2 sind die Werte

Tabelle 2. Bildungsenthalpie der III–V-*Verbindungen bei* $T = 298°\mathrm{K}$. $A^{\mathrm{III}}{}_{\mathrm{fest}} + B^{\mathrm{V}}{}_{\mathrm{fest}} \rightarrow A^{\mathrm{III}}B^{\mathrm{V}}{}_{\mathrm{fest}}$

Nr.	$A^{\mathrm{III}}B^{\mathrm{V}}$	$\Delta W^{298}_{\mathrm{Bildung}}$ (kcal/mol)	
1	InSb	− 6,94 ±0,22	(Ref. 3)
2		− 7,8 ±0,3	(Ref. 6)
3		− 7,3	Tab. 1 Wert 1 und Wärmeinhalte von In und Sb (Ref. 1) ($C_{\mathrm{InSb}} = 12$ gesetzt)
4	GaSb	− 9,94 ±0,44	(Ref. 3)
5		−13,3	Tab. 1 Wert 2 und Wärmeinhalte von Ga, Sb nach Ref. 1. ($C_{\mathrm{GaSb}} = 12$)
6	InAs	−14,8 ±1,3	(Ref. 3)
7		−13,5	Tab. 1 Wert 7 und Wärmeinhalte von In und As nach Ref. 1. ($C_{\mathrm{InAs}} = 12$)
8	GaAs	−25,8	Tab. 1 Wert 8 und Wärmeinhalte von Ga und As nach Ref. 1. ($C_{\mathrm{GaAs}} = 12$)
9		−12,3	Tab. 1 Wert 9 und Wärmeinhalte von Ga und As nach Ref. 1. ($C_{\mathrm{GaAs}} = 12$)
10	InP	−22,0	Tab. 1 Wert 12 und thermochem. Daten (Ref. 9)
11		−22,1	Tab. 1 Wert 10 berechnet nach $W^{298°\mathrm{K}}_{\mathrm{Bild.}} = W^{1000°\mathrm{K}}_{\mathrm{Bild.}} - C_{\mathrm{InP}}\Delta T + (W^{1000} - W^{298})_{\mathrm{In}} + (W^{1000} - W^{298})_{\mathrm{P_2}}$ ($C_{\mathrm{InP}} = 12$, W^T−Werte nach Ref. 1)
12		−21,6	Tab. 1 Wert 11 wie vorstehend berechnet.
13	InSb	− 6,8 ±0,6	(Ref. 15)

2.2 *Bildungsenthalpie bei* $T = 298°\mathrm{K}$

Dieser Bildungsenthalpie liegt folgende Reaktion zugrunde:

$$A^{\mathrm{III}}_{\mathrm{fest}} + B^{\mathrm{V}}_{\mathrm{fest}} \xrightarrow{298°\mathrm{K}} A^{\mathrm{III}}B^{\mathrm{V}}_{\mathrm{fest}} + \Delta W^{298°\mathrm{K}}_{\mathrm{Bild.}}$$

Direkt bestimmt wurde dieser Wert für InSb, GaSb und InAs von SCHOTTKY und BEVER[3] mit zusammengefasst mit Quelle bzw. Berechnungsweise.

2.3 *Bildungsenthalpie aus gasförmigen Elementen bei* 298°K

FOLBERTH[9] berechnete die Enthalpie folgender Reaktion:

$$A^{\mathrm{III}}_{\mathrm{gas}} + B^{\mathrm{V}}_{\mathrm{gas}} \xrightarrow{298°\mathrm{K}} A^{\mathrm{III}}B^{\mathrm{V}}_{\mathrm{fest}}$$

Tabelle 3. Bildungsenthalpie für die Reaktion $A^{\mathrm{III}}_{\mathrm{gas}} + B^{\mathrm{V}}_{\mathrm{gas}} \rightarrow A^{\mathrm{III}}B^{\mathrm{V}}_{\mathrm{fest}}$ *bei* 298°K *nach* Ref. 9. *sowie Reaktionsentropie und freie Bildungsenergie*

Nr.	$A^{\mathrm{III}}B^{\mathrm{V}}$	$\Delta W^{298°\mathrm{K}}_{0\,\mathrm{Bild.}}$ (kcal/mol)	$\Delta S^{298°\mathrm{K}}_{0\,\mathrm{Bild.}}$ (cal/molgrad)	$\Delta F^{298°\mathrm{K}}_{0\,\mathrm{Bild.}}$ (kcal/mol)	Bemerkungen
1	InSb	−128,1	−63,7	−109,1	
2	GaSb	−138,4	−66,1	−118,7	Partialdruck der Gase = 1 atm
3	InAs	−130,4	−67,3	−110,3	
4	GaAs	−140,9	−62,1	−122,4	
5	InP	−154,6	−69,0	−134,0	
6	AlSb	−150,9	−68,3	−130,5	

Diese Werte unterscheiden sich von denen in Tabelle 2 um die Gitterenergie der III–V-Elemente (siehe Tabelle 3). (Mit den Werten der Tabelle 4 lassen sich die $\Delta S^{298}_{\text{Bild.}}$—und $\Delta F^{298}_{\text{Bild.}}$—Werte für diese Reaktion gewinnen. Sie sind mit aufgeführt).

3. ENTROPIE

3.1 *Bildungsentropie*

Es liegen Werte für InSb, GaSb (bei 298°K) sowie InAs, GaSb, GaAs und InP (bei 900 bzw. 1000°K) vor. Tabelle 4.

3.2 *Absolutentropie*

Unter Einbeziehung der Absolutentropiewerte der Elemente[1] bei den entsprechenden Temperaturen lassen sich aus den Werten der Tabelle 4 die Absolutentropien der III–V-Verbindungen berechnen (Tabelle 5). Unter Berücksichtigung des Faktors

$$\Delta S = S^T - S^{298} = C \ln \frac{T}{298}$$

lassen sich daraus die Werte bei 298°K gewinnen, wenn man $C = \text{konst} \approx 12$ annimmt (Tabelle 6). Für anorganische Verbindungen kann man unbekannte Entropiewerte näherungsweise mit der Eastman'schen Näherungsformel berechnen:

$$S_{\text{Atom}} = 1{,}5R \ln A + R \ln V - 1{,}5R \ln T + 12{,}5$$

S_{Atom} = mittlere Atomentropie
V = mittl. Atomvolumen
A = mittleres Atomgewicht
T = Schmelzpunkt °K

Versucht man, diese Formel auf die III–V-Verbindungen anzuwenden, so zeigt sich, dass verglichen mit zuverlässigen bekannten Werten einiger III–V-Verbindungen die Beträge um etwa 5 cal/molgrad zu hoch ausfallen. Unter Berücksichtigung dieser Korrektur sind die so ermittelten Werte in Tabelle 6 mit aufgeführt. Bei der Berechnung wurde von den Röntgendichten Gebrauch gemacht, wie sie sich aus den Gitterkonstanten[13] ergeben:

InSb = 5,71
GaSb = 5,56
AlSb = 4,24
InAs = 5,58
GaAs = 5,24
InP = 4,74
GaP = 4,07

Tabelle 4. Bildungsentropie der III–V-*Verbindungen*

Nr.	$A^{III}B^{V}$	Temp. (°K)	$\Delta S^{T}_{\text{Bild.}}$ (cal/molgrad)	
1	InSb	298	$-3{,}98 \pm 0{,}52$	In fest; Sb fest (Ref. 3)
2	GaSb	298	$-2{,}8 \pm 0{,}6$	Ga fest; Sb fest (Ref. 3)
3		900	$-30{,}7$	Ga flüss.; Sb_2 Gas (1 atm) (Ref. 4)
4		900	$-17{,}7$	Ga flüss.; Sb_4 Gas (1 atm) (Ref. 4)
5	InAs	1000	$-27{,}7$	In flüss.; As_2 Gas (1 atm) (Ref. 4)
6		1000	$-17{,}8$	In flüss.; As_4 Gas (1 atm) (Ref. 4)
7	GaAs	1080	$-38{,}0$	Ga flüss.; (As_2, As_4) Gas (1 atm) (Ref. 5)

Tabelle 5. Entropie der III–V-*Verbindungen bei Temperaturen* > 298°K

Nr.	$A^{III}B^{V}$	Temp. (°K)	S (cal/molgrad)	
1	GaSb	900	22	(Ref. 4)
2		900	30	(Ref. 4)
3	InAs	1000	30,25	(Ref. 4)
4		1000	30,9	(Ref. 4)
5	GaAs	1080	$34{,}4 \pm 1$	(Ref. 5)
6	InP	1000	25 ± 1	(Ref. 5)

4. FREIE ENERGIE

4.1 *Freie Bildungsenergie bei* $T > 298$°K

Bei höheren Temperaturen haben wir es im Falle der III–V-Verbindungen im allgemeinen mit folgender Reaktion zu tun:

$$A^{III}B^{V}_{\text{fest}} \rightarrow A^{III}_{\text{flüss.}} + \frac{1}{x} B^{V}_{x\text{gasf.}}$$

Tabelle 6. *Entropie der* III–V-*Verbindungen bei* 298°K

Nr.	$A^{III}B^{V}$	S (cal/molgrad)	
1	InSb	20,84	(Ref. 3)
2	GaSb	17,9	Aus Tab. 4 Nr. 2 und $S_{Ga} = 10{,}2$; $S_{Sb} = 10{,}5$ (Ref. 1) berechnet.
3		16,8	Aus Tab. 5 Nr. 2 abzüglich $C \ln(900/298) = 13{,}2$ ($C = 12$)
4	InAs	15,8	Aus Tab. 5 Nr. 3/4 abzgl. $C \ln(1000/298) = 14{,}8$ ($C = 12$)
5	GaAs	19,9	Aus Tab. 5 Nr. 5 abzgl. $C \ln(1080/298) = 15{,}5$ ($C = 12$)
6	InP	11,2	Aus Tab. 5 Nr. 6 abzgl. $C \ln(1000/298) = 13{,}8$ ($C = 11{,}5$)
7	InSb	20,5	
8	GaSb	17,4	
9	AlSb	14,1	Eastman'sche Näherungsformel
10	InAs	15,9	abzüglich 5 cal/molgrad
11	GaAs	12,2	
12	InP	13,4	
13	GaP	9,3	

Die Änderung der freien Energie hängt mit ΔW und ΔS nach

$$\Delta F^T = \Delta W^T - T\Delta S^T$$

zusammen. Ausserdem gilt für obige Reaktion:

$$\Delta F^T = \Delta F_0{}^T + RT \ln p_{B_x}{}^{(1/x)},$$

worin $\Delta F_0{}^T$ die Änderung der freien Energie bei einem Partialdruck $p_{B_x} = 1$ atm darstellt. Im Falle des Gleichgewichts ($\Delta F_T = 0$) ist

$$\Delta F_0{}^T = \frac{RT}{x} \ln p_{B_x},$$

(p_{B_x} = Gleichgewichtsdampfdruck)

und gibt $\Delta F_0{}^T$ unmittelbar den Gleichgewichtsdampfdruck. Eine Zusammenstellung von $\Delta F_0{}^T$ findet sich in Tabelle 7.

4.2 *Freie Bildungsenergie bei* 298°K

Aus den $\Delta W_0{}^{298°K}$—und $\Delta S_0{}^{298°K}$—Werten

Tabelle 7. *Freie Bildungsenergie der* III–V-*Verbindungen bei Temperaturen* > 298°K

Nr.	$A^{III}B^{V}$	Temp. (°K)	$\Delta F^T_{0\,\text{Bild.}}$ (k cal/mol)	
1	InSb	798	$-3{,}62 \pm 0{,}40$	In flüss., Sb fest (Ref. 5)
2		930	$-2{,}8 \pm 1$	In flüss., Sb_2 Gas (Ref. 5)
3	GaSb	900	$-13{,}4$	Aus Tab. 1 Nr. 2 und Tab. 4 Nr. 3 berechnet. Ga flüss. Sb_2 Gas
4		900	$-9{,}6$	Aus Tab. 1 Nr. 3 und Tab. 4 Nr. 4 berechnet, Ga flüss. Sb_4 Gas
5	InAs	1000	$-15{,}4$	Aus Tab. 1 Nr. 4 und Tab. 4 Nr. 5 In flüss., As_2 Gas
6		~ 1000	$-11{,}3$	Aus Tab. 1 Nr. 6 und Tab. 4 Nr. 5 In flüss., As_2 Gas
7		1000	$-8{,}2$	Aus Tab. 1 Nr. 5 und Tab. 4 Nr. 6 In flüss., As_4 Gas
8		~ 1000	$-4{,}2$	Aus Tab. 1 Nr. 7 und Tab. 4 Nr. 6 In flüss., As_4 Gas
9	GaAs	1080	$-15{,}3 \pm 0{,}5$	Ga flüss., As_2 Gas (Ref. 5)
10		1080	$-8{,}8$	Aus Tab. 1 Nr. 9 Tab. 5 Nr. 2 und S_{Ga}, S_{As_4} bei 1080°K Ga flüss., As_4 Gas
11	InP	1000	$-9{,}2 \pm 0{,}5$	In flüss., P_2 Gas (Ref. 5)

sind die $\Delta F_0{}^{298}$—Werte der Tabelle 8 ermittelt worden.

5. DAMPFDRUCK DER WICHTIGSTEN III-V-VERBINDUNGEN

5.1 As-*Dampfdruck über* InAs

VAN DEN BOOMGAARD und SCHOL[11] bestimmte

Tabelle 8. Freie Bildungsenergie der III–V-*Verbindungen bei* $T = 298°K$ *für die Reaktion* $A^{III}{}_{fest} + B^{V}{}_{fest} \rightarrow A^{III}B^{V}{}_{fest}$

Nr.	$A^{III}B^{V}$	$\Delta F^{298}_{0\,Bild.}$ (k cal/mol)	
1	InSb	− 6,04 ± 0,6	(Ref. 3)
2		− 5,76 ± 0,36	(Ref. 3)
3		− 6,56	Aus Tab. 2 Nr. 2 und Tab. 6 Nr. 1 und Entropiewerte der Elemente
4		− 6,06	Aus Tab. 2 Nr. 3 und Tab. 6 Nr. 1 und Entropiewerte der Elemente
5	GaSb	− 8,96	Aus II, 4 und VI, 2/3 und Entropiewerte der Elemente
6		−12,3	Aus II, 5 und VI, 2/3 und Entropiewerte der Elemente
7	InAs	−12,7	Aus II, 6 und VI, 4 und Entropiewerte der Elemente
8		−11,4	Aus II, 7 und VI, 4 und Entropiewerte der Elemente
9	GaAs	−26,2	Aus II, 8 und VI, 5 und Entropiewerte der Elemente
10		−12,7	Aus II, 9 und VI, 5 und Entropiewerte der Elemente
11	InP	−18,0	Aus II, 10 und VI, 6 und Entropiewerte der Elemente
12		−18,1	Aus II, 11 und VI, 6 und Entropiewerte der Elemente
13		−17,6	Aus II, 12 und VI, 6 und Entropiewerte der Elemente

den As_x-Dampfdruck direkt für den Temperaturbereich von 1150°K bis zum Schmelzpkt. Die Ergebnisse sind in Abb. 1 eingezeichnet. GOLDFINGER und JEUNEHOMME[4] bestimmen den As_x-Druck über InAs für den Temperaturbereich 911–1159°K massenspektrometrisch, indem sie die Intensität des Molekülionenstroms As_x^+ nach entsprechender Eichung direkt in p umrechnen. Aus ihren Daten errechnen sich für den betrachteten T-Bereich die Gleichungen:

$$\log p_{As_4} = -\frac{22\,860}{T} + 15{,}65 \qquad (1)$$

$$\log p_{As_2} = -\frac{18\,900}{T} + 12{,}2 \qquad (2)$$

wonach bei tieferen Temperaturen As_2 überwiegt, bei höheren Temperaturen As_4. Es handelt sich also um das gekoppelte Gleichgewicht:

$$\text{In}_{\text{flüss.}} + \tfrac{1}{4}(1-\alpha)As_4 \rightleftharpoons \tfrac{1}{2}\alpha As_2 + \text{In}_{\text{flüss.}} \qquad (3)$$

(InAs)

Wie aus Abb. 1 ersichtlich führen die Gleichungen (1) und (2) selbst unter Berücksichtigung der Tatsache, dass eine geradlinige Extrapolation bis zum Schmelzpunkt nicht zulässig ist, zu sehr niedrigen Werten. Dies legt den Gedanken nahe, dass es sich nicht um Gleichgewichtswerte, sondern um „freie Verdampfung" handelt. Eine

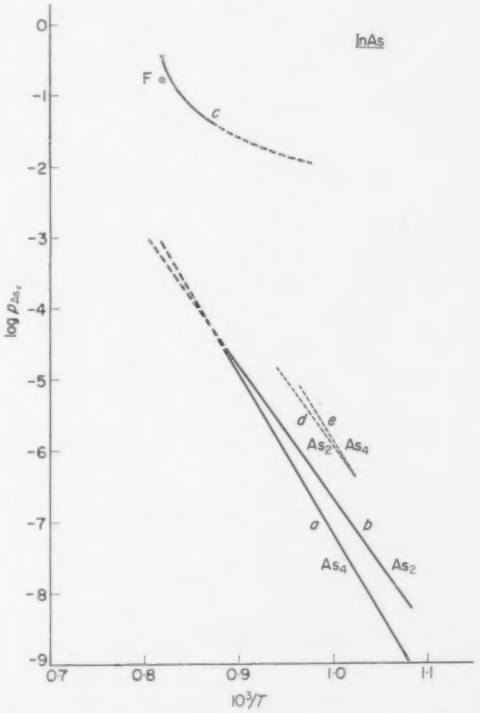


ABB. 1. As-Dampfdruck über InAs. a, b: Partialdrucke nach Ref. 4. c: As-Druck nach Ref. 11. d, e: Partialdrucke aus therm Daten berechnet. f: Dampfdruck beim Smp. nach Ref. 8.

Bestätigung erfährt diese Auffssung durch die Ergebnisse von GUTBIER[14], der das gleiche System ebenfalls massenspektrometrisch untersuchte und ein wesentlich grösseres Verhältnis As_4/As_2 fand und ausserdem feststellte, dass dieses Verhältnis im Laufe der Verdampfung von As_x nicht konstant war.

Man kann versuchen, dan Dampfdruckverlauf bei z.B. 1000°K aus bekannten rein thermisch bestimmten Grössen zu berechnen, wobei von folgender Gleichung Gebrauch zu machen wäre:

$$\Delta F_0^T = -RT \ln p_{As_x}^{1/x} = \Delta W^T - T\Delta S_0^T \qquad (4)$$

Für $\Delta W_{\text{Reakt.}}^{298°K}$ gilt:

$$InAs \xrightarrow{298°K} (In_{fest} + As_{fest}) \xrightarrow{298°K} In_{flüss.}^{298°} + \tfrac{1}{4}As_{gas}^{298°K}$$

$$\Delta W_{\text{Bild. InAs}} = 14{,}8 \text{ kcal/mol};^{(3)}$$

$$\Delta W_{\text{Schmels In}} = 0{,}75 - (C_{\text{fest}} - C_{\text{fl.}})130 =$$

$$= 0{,}715 \text{ kcal/mol}$$

$$W_{\text{verd. As}_4} = \tfrac{1}{4} \times 30{,}4 \text{ kcal/mol}^{(2)}$$

Daraus:

$$\Delta W_{\text{Reakt.}}^{298°K} = 23{,}1 \text{ kcal/mol}$$

Bei 1000°K:

$$\Delta W^{1000°} = 23{,}1 + (W_{In}^{1000°} - W_{In}^{298°}) +$$

$$+ \tfrac{1}{4}(W_{As_4}^{1000°} - W_{As_4}^{298°}) - C_{InAs}\Delta T;$$

$$W_{In}^{1000°} - W_{In}^{298°} = 5{,}72^{(1)} \text{ kcal/mol}$$

$$\tfrac{1}{4}(W_{As_4}^{1000°} - W_{As_4}^{298°}) = 3{,}4^{(1)} \text{ kcal/mol}$$

$$\Delta W^{1000°} = 23{,}7 \text{ kcal/mol}$$

$$\Delta S_0^{1000°} = S_{In}^{1000°} + \tfrac{1}{4}S_{As4} - \left(C_{InAs} \ln \frac{1000}{298} + 15{,}8\right)$$

$$S_{In}^{1000} = 24{,}8^{(1)} \text{cal/molgrad}$$

$$\tfrac{1}{4}S_{As_4}^{1000} = 22{,}8^{(1)} \text{ cal/molgrad}$$

$$C_{InAs} = 12 \text{ cal/molgrad}$$

$$\Delta S_0^{1000} = 17{,}0 \text{ cal/molgrad}$$

Damit errechnet sich aus (4) für

$$p_{As_4} = 1{,}25 \times 10^{-6} \text{ atm bei } 1000°K$$

Analog für:

$$p_{As_2} = 1{,}15 \times 10^{-6} \text{ atm bei } 1000°K$$

Die Werte sind in Abb. 1 eingetragen.

5.2 As-*Dampfdruck über* GaAs

In Abb. 2 sind die Werte von VAN DEN BOOMGAARD und SCHOL,[11] sowie die Partialdrucke und

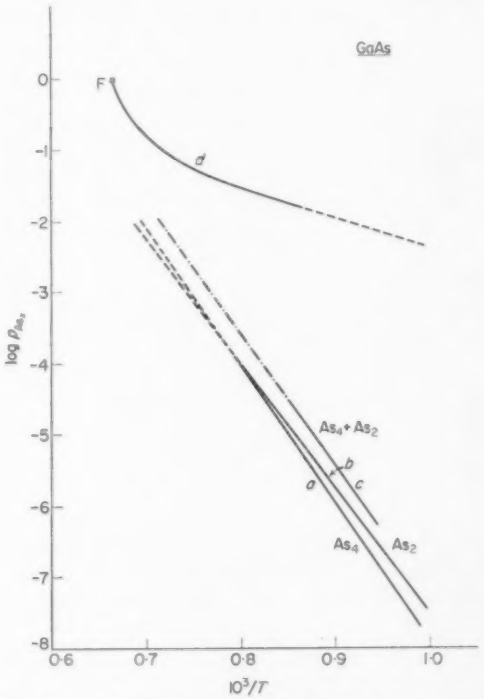


ABB. 2. As-Dampfdruck über GaAs. a, b, c: Partialdrucke und Gesamtdrucke nach Ref. 5. d: As-Dampfdruck nach Ref. 11. f: As:Dampfdruck beim Smp. nach Ref. 8.

der Gesamtdruck nach DROWART und GOLDFINGER[5] für den T-Bereich 950–1200°K eingezeichnet. Aus den Daten der letzteren Autoren errechnen sich die Gleichungen für die Partialdrucke zu:

$$\log p_{As_4} = -\frac{19\,320}{T} + 11{,}44 \qquad (5)$$

$$\log p_{As_2} = -\frac{17\,340}{T} + 9{,}86 \qquad (6)$$

Die Verhältnisse sind ähnlich wie im Falle des InAs.

Die Berechnung aus thermischen Grössen ist im Falle des GaAs nicht sinnvoll, da vor allem die Bildungsenthalpie nur sehr ungenau bekannt ist.

5.3 P-*Dampfdruck über* InP

Die Daten von VAN DEN BOOMGAARD und SCHOL,[11] von DROWART und GOLDFINGER[5] im T-Bereich 775–950°K und von WEISER[10] ergeben die Kurven der Abb. 3. Für die Partialdrucke errechnen sich nach (5) die Gleichungen:

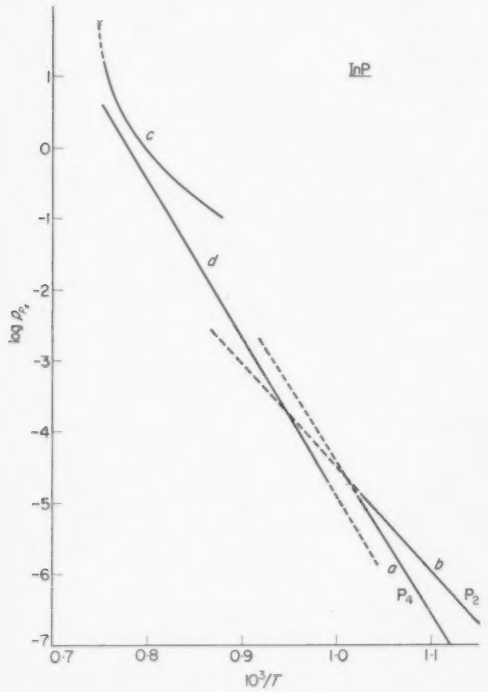


ABB. 3. P-Dampfdruck über InP. a, b: Partialdrucke nach Ref. 5. c: Gesamtdruck nach Ref. 11. d: P_4-Druck nach Ref. 10.

$$\log p_{P_4} = -\frac{21\,300}{T} + 16{,}84 \qquad (7)$$

$$\log p_{P_2} = -\frac{14\,850}{T} + 10{,}32 \qquad (8)$$

Die Berechnung aus thermochemischen Daten, wie im Falle des InAs führt für den P_4-Druck zu der Gleichung:

$$\log p_{P_4} = -\frac{22\,600}{T} + 19{,}4 \qquad (9)$$

die etwas zu hohe Werte ergibt.

5.4 Sb-*Dampfdruck über* InSb *und* GaSb

Werte für den (Sb_2, Sb_4)-Dampfdruck über InSb und GaSb geben GOLDFINGER und JEUNEHOMME[4] bzw. DROWART und GOLDFINGER[5].

LITERATUR

1. L. L. QUILL, *The Chemistry and Metallurgy of Miscellaneous Materials*, p. 13. McGraw-Hill, New York (1950).
2. D'ANS-LAX, *Taschenbuch für Chemiker und Physiker*, p. 312. Springer-Verlag, Berlin (1949).
3. W. F. SCHOTTKY und M. B. BEVER, *Acta met.* **6,** 320 (1958).
4. P. GOLDFINGER and M. JEUNEHOMME, *Joint Conference on Mass Spectrometry, September*, 1958 (*London*). Pergamon Press, London (1959).
5. J. VON DROWART und P. GOLDFINGER, *J. Chim. Phys.* **55,** 721 (1958).
6. O. KUBASCHEWSKY und E. L. EVANS, *Metallurgical Thermochemistry*. Pergamon Press, London (1956).
7. O. J. KLEPPA, *J. Amer. Chem. Soc.* **77,** 897 (1955).
8. O. G. FOLBERTH, *J. Phys. Chem. Solids* **7,** 295 (1958).
9. O. G. FOLBERTH, *Unveröffentlichte Arbeit*.
10. K. VON WEISER, *J. Phys. Chem.* **61,** 513 (1957).
11. J. VAN DEN BOOMGAARD und K. SCHOL, *Philips Res. Rep.* **12,** 127 (1957).
12. T. S. LIU und E. A. PERETTI, *Trans. Amer. Soc. Metals* **45,** 677 (1953).
13. G. GIESECKE und H. PFISTER, *Acta Cryst.* **11,** 369 (1958).
14. H. GUTBIER, *Z. Naturf.* **14**a, 32 (1959).
15. A. SCHNEIDER und H. KLOTZ, *Naturw.* **46,** 141 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 46–53. Printed in Great Britain

QUANTITATIVE PHOTOVOLTAIC EVALUATION OF THE RESISTIVITY HOMOGENEITY OF GERMANIUM SINGLE CRYSTALS

J. OROSHNIK AND A. MANY*

Sylvania Research Laboratories, Bayside, N.Y.

(*Received* 20 *July* 1959)

Abstract—An apparatus, based on the bulk photovoltaic effect, is described by means of which quantitative estimates of the resistivity changes throughout germanium single-crystal samples may be realized. The experimental technique consists of scanning the sample with a light spot and recording the photovoltage, or integrated photovoltage, versus position of the light spot. By taking into account the magnitudes of the photovoltages and photoconductance upon illumination, the resistivity changes in Ω-cm can be derived from the scans of integrated-photovoltage versus position. For germanium of 10 Ω-cm resistivity or less, the changes in resistivity determined by the use of this system are usually within 30 per cent of the actual changes determined with the two-point probe. For material that is homogeneous in lifetime, the error may be as little as 3 per cent. With this technique, single-crystal wafers may also be explored so as to determine sections of best uniformity. The method is essentially non-destructive.

Résumé—On décrit un appareil basé sur l'effet photovoltaïque interne qui permet d'estimer quantitativement les variations de résistivité dans des spécimens monocristallins de germanium. La technique expérimentale consiste à balayer le spéciment par un point lumineux, et à enregistrer la tension photovoltaïque ou la tension photovoltaïque intégrée en fonction de la position du point lumineux. En tenant compte de l'amplitude des photovoltages et de la photoconductance en présence de lumière les variations de résistivité peuvent être déduites des courbes de photovoltage intégré en fonction de la position. Pour du germanium de résistivité 10 Ω-cm au plus les changements de résistivité déterminés par ce système sont normalement en accord à mieux que 30 pour cent avec les changements de résistivité déterminés exactement par une méthode à deux sondes. Si la durée de vie du spécimen est uniforme, l'erreur peut être aussi faible que 3 pour cent. Par cette technique, des tranches monocristallines peuvent être explorées pour y trouver les régions les plus uniformes. La méthode est essentiellement non-destructive.

Zusammenfassung—Es wird ein auf dem Photospannungseffekt im Volumen basierender Apparat beschrieben, mit dessen Hilfe quantitative Schätzungen der Widerstandsänderungen in Proben von Germanium-Einkristallen verwirklicht werden können. Die experimentelle Technik besteht darin, die Probe mit einem Lichtfleck abzutasten und die Photospannung oder integrierte Photospannung als Funktion des Ortes des Lichtflecks zu registrieren. Wenn man die Grössen der Photospannungen und Photoleitfähigkeit bezüglich der Lichtstärke in Rechnung stellt, können die Widerstandsänderungen in Ω cm aus der integrierten Photospannung als Funktion des Ortes hergeleitet werden. Für Germanium von 10 Ω cm Widerstand oder weniger liegen die durch Anwendung dieses Gerätes bestimmten Widerstandsänderungen gewöhnlich innerhalb von 30 Prozent der tatsächlichen Änderungen, wie sie durch die Zweispitzenmethode bestimmt werden. Für Material mit homogener Lebensdauer beträgt der Fehler nur 3 Prozent. Mit dieser Technik können auch Einkristall-Scheiben untersucht werden, um dadurch Teile bester Gleichmässigkeit zu bestimmen. Die Methode ist im wesentlichen zerstörungsfrei.

* Present address: Department of Physics, The Hebrew University, Jerusalem, Israel.

INTRODUCTION

IT HAS been shown in a previous paper,[1] that the "bulk photovoltaic effect" can be used to explore the resistivity homogeneity of germanium single crystals. Very detailed pictures of the resistivity gradients throughout the bulk of both filaments (rectangular bars) and wafers can be realized by means of this effect, and consequently, the resistivity topography of the sample can be qualitatively estimated. It is the purpose of this paper to describe an apparatus that will give *quantitative* estimates of the resistivity changes throughout a crystal sample.

The photovoltaic scanning method as herein described is both fast and sensitive. A complete scan of the direct and integrated photovoltage signals can be made within one minute. A change in resistivity of 0·1 per cent in 10 Ω-cm material over several millimeters usually yields distinct and measurable photosignals. The photovoltage scan is a continuous measurement over a sample in contrast to the two-point probe measurement which is, in general, a point-to-point scan.

THEORY

A rudimentary explanation of the mechanism of the bulk photovoltaic effect can be appreciated with the aid of Fig. 1. Part (a) indicates an *n*-type crystal having a resistivity profile as shown. We assume perfectly ohmic contacts at the ends. The electron-energy-band picture for such a crystal is shown in part (b); E_c, E_f and E_v have the usual meaning of bottom of the conduction band, Fermi level and top of the valence band. Suppose now that this crystal is illuminated at point x_3. The excess electrons generated by the light will tend to drift to the left—downward to a lower electron-energy region. The excess holes will tend to drift to the right. Thus a separation of charge takes place, and the galvanometer indicates a positive voltage. A similar argument will show that a negative photovoltage will be generated on illumination at point x_1.

Zero photovoltage signals will correspond to maxima and minima in the resistivity profile. If the crystal is illuminated at point x_2 where the resistivity and energy minima occur, excess electrons will tend to remain at that point, and excess holes to drift to either side of point x_2. Thus a zero photovoltage will be observed. In the case of

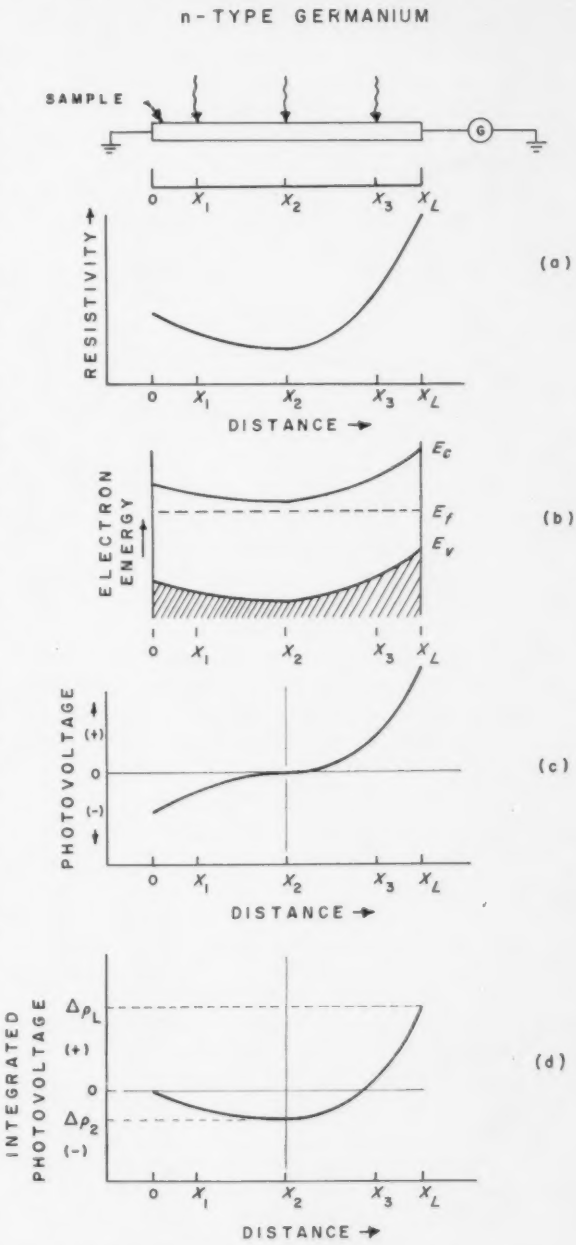


FIG. 1. Correlation of photovoltage with resistivity and electron energy.

p-type material reasoning similar to the foregoing holds, except that the voltages are reversed. A

perfectly homogeneous crystal will show no bulk photovoltaic effect of this type. One should bear in mind that resistivity variations constitute small, extended junctions throughout the bulk of the crystal. There is no fundamental difference in the mechanism of photovoltage generation between these small n–n^+ (or p–p^+) junctions and, for example, a p–n junction. Photovoltages due to resistivity gradients are usually of the order of hundreds of microvolts, while p–n junctions yield signals in the range of tenths of a volt.

The complete photovoltage scan for this particular crystal appears in part (c) of Fig. 1. Because the photovoltage depends on the existence of a resistivity gradient, it is a derivative quantity. If the photovoltage signal is integrated along the direction of scan (electronically or otherwise), the integrated signal will correspond qualitatively to the *changes* in resistivity. Part (d) of Fig. 1 shows the integrated photosignal of part (c).

A general treatment of the bulk photovoltaic effect has been carried out by TAUC[2]. In this paper the theory will be developed in a form which is directly applicable to the experimental technique described here. The present treatment rests on a number of simplifying assumptions which can usually be fairly well satisfied in practice. These assumptions are:

(a) The resistivity is constant throughout any cross section, but may vary along the length (taken as the x-direction) of the filament. The geometric cross section is uniform.

(b) The end contacts are ohmic. This implies that the concentration of excess carriers Δp at the contacts is always zero, and that no photovoltage is generated owing to contact barriers.

(c) The filament effective lifetime[3] is uniform throughout the length of the filament.

(d) A narrow strip of the sample is illuminated, the width of which is small compared with a diffusion length.

(e) The illumination is at a low enough level so that $\Delta p \ll (p_0+n_0)$, where p_0 and n_0 are the equilibrium hole and electron concentrations.

Under the conditions of measurement, the current through the filament may be taken as zero. By expressing this current in terms of the hole and electron drift and diffusion currents, one may obtain an expression for the internal field E at any point along the filament. The photovoltage U is then given by integrating E between the two ends, $x = 0$ and $x = L$, of the filament, i.e.

$$U = \int_0^L E\,dx$$

$$= \frac{kT}{q}\int_0^L \frac{\mu_p\frac{dp_0}{dx} - \mu_n\frac{dn_0}{dx} + (\mu_p-\mu_n)\frac{d}{dx}\Delta p}{\mu_p p_0+\mu_n n_0+(\mu_p+\mu_n)\Delta p}\,dx \tag{1}$$

where μ_p and μ_n denote the hole and electron mobility, respectively. In deriving equation (1), the Einstein relationship between carrier mobility and diffusion constant was used.

Assumption (e) allows the development of the expression in the denominator of the integrand into a power series, with the neglect of second and higher order terms. By using the relationship $p_0 n_0 =$ a constant, or $p_0(dn_0/dx)+n_0(dp_0/dx) = 0$, and the fact that $U = 0$ in the dark ($\Delta p = 0$), we can reduce equation (1) to:

$$U = \frac{kT}{q}\int_0^L \frac{1}{n_0}\frac{dn_0}{dx}\rho_0\Delta\sigma\,dx - \frac{kT}{q}\left(\frac{\mu_n-\mu_p}{\mu_n+\mu_p}\right)\int_0^L \rho_0\left[\frac{d}{dx}\Delta\sigma\right]dx \tag{2}$$

where

$$1/\rho_0 \equiv \sigma_0 \equiv q(\mu_n n_0+\mu_p p_0)$$

and

$$\Delta\sigma = q(\mu_n+\mu_p)\Delta p \tag{3}$$

The second integral on the right-hand side of equation (2) can be integrated by parts, leading to

$$U = \frac{kT}{q}\left\{\int_0^L \frac{1}{n_0}\frac{dn_0}{dx}\rho_0\Delta\sigma\,dx + \left(\frac{\mu_n-\mu_p}{\mu_n+\mu_p}\right)\int_0^L \Delta\sigma\frac{d\rho_0}{dx}\,dx\right\} \tag{4}$$

since $\Delta p = 0$ at the ends, $x = 0$ and $x = L$, of the filament—assumption (b). For extrinsic n-type

material, such that $p_0 \ll n_0$ at any point x,

$$U = -\frac{kT}{q}\left(\frac{2}{1+\mu_n/\mu_p}\right)\int_0^L \Delta\sigma(d\rho_0/dx)\,dx \qquad (5)$$

$$\approx -\frac{kT}{q}\left(\frac{2}{1+\mu_n/\mu_p}\right)\left(\frac{d\rho_0}{dx}\right)\int_0^L \Delta\sigma\,dx$$

The approximate equality in equation (5) follows from the fact that $\Delta\sigma$ differs from zero only in a narrow region that extends to the order of a diffusion length on either side of the illuminated strip. In this narrow region, $d\rho_0/dx$ may be assumed to be essentially constant. The integration of both sides of equation (5) over a cross section yields,

$$U \cdot A = -\frac{2kT}{q}\left(\frac{1}{1+\mu_n/\mu_p}\right)\frac{d\rho_0}{dx}\int_0^A\int_0^L \Delta\sigma\,dx\,dA \qquad (6)$$

$$= -\frac{2kT}{q}\left(\frac{1}{1+\mu_n/\mu_p}\right)\frac{d\rho_0}{dx}L^2(\Delta G)$$

where ΔG is the excess conductance of the filament due to the illuminated strip, and A is the cross-sectional area. The second equality follows[4] from assumptions (a) and (e), i.e. the cross section is uniform, and $\Delta p \ll p_0+n_0$. Hence, for an n-type sample under consideration,

$$\frac{d\rho_0}{dx} = -\frac{q}{2kT}\left(1+\frac{\mu_n}{\mu_p}\right)\frac{A}{L^2}\left(\frac{1}{\Delta G}\right)U \qquad (n\text{-type}) \qquad (7)$$

By a similar derivation, one can show that for a p-type sample,

$$\frac{d\rho_0}{dx} = +\frac{q}{2kT}\left(\frac{1+\mu_n/\mu_p}{\mu_n/\mu_p}\right)\frac{A}{L^2}\left(\frac{1}{\Delta G}\right)U \qquad (p\text{-type}) \qquad (8)$$

From assumption (c) it follows that, for a constant light intensity, ΔG is independent of the position of the illuminated strip. Hence, the resistivity gradient at any point along the filament is directly proportional to the photovoltage generated when the light strip is at that point. Once ΔG has been measured, the variation of the resistivity *gradient* can easily be evaluated from the photovoltage scan. Moreover, if the photovoltage is integrated during the scan then the variation of the *resistivity* along the filament is obtained directly.

EXPERIMENTAL APPARATUS

The experimental set-up and associated circuitry is shown schematically in Fig. 2. The mechanical arrangement of the various parts is shown in part (a). Part (b) shows the circuitry for the x-axis (light-spot position) signal to the recorder.* The zero-set bias permits choosing any location on the crystal as the zero position, and having it correspond to a predetermined zero position on the x-axis of the recorder. S3 and S4 are input-shorting switches, allowing range, span and zero-set adjustments to be made on the recorder. Sl is a polarity reversing switch.

Part (c) of Fig. 2 is a ten-pole, five-position switching system. In position No. 1, the direct photovoltage signal is recorded. Position No. 2 is a shorting position; the crystal, input and output of the integrator, and the d.c. voltmeter-amplifier† input, are all simultaneously shorted. Position No. 2 also provides for the gain (range and span) setting of the recorder. The calibrating circuit is adjusted for 1 V output which is measured by the voltmeter-amplifier set on the 1 V, full-scale range. The amplifier in turn delivers 1 V† to the recorder, the gain of which is adjustable. When S7 is open, the amplifier input is essentially shorted.

The integration of the photovoltage signal is carried through in position No. 3. The integrating circuit is composed of a resistor and capacitor in series, having a time-constant large compared with the scanning time. The scanning time is several seconds.

Positions No. 4 and No. 5 enable the measurement of ΔG. For low-level illumination $\Delta G \approx -\Delta R/R_0^2$, where R_0 is the filament resistance in darkness, and $\Delta R \equiv R_0-R$, the change in resistance upon illumination. This is a simple, Wheatstone bridge measurement. The bridge elements are arranged in such a way that R_0 and R are read directly from the helipots $R3$, $R4$ and $R5$. The bridge is balanced (position No. 5) by using the d.c. voltmeter as a null indicator. If there is (in

* Model 2A x–y recorder, F. L. Moseley Co., Pasadena, California.

† Model 203 d.c. microvolt-ammeter and amplifier, Kin Tel Co., San Diego, California.

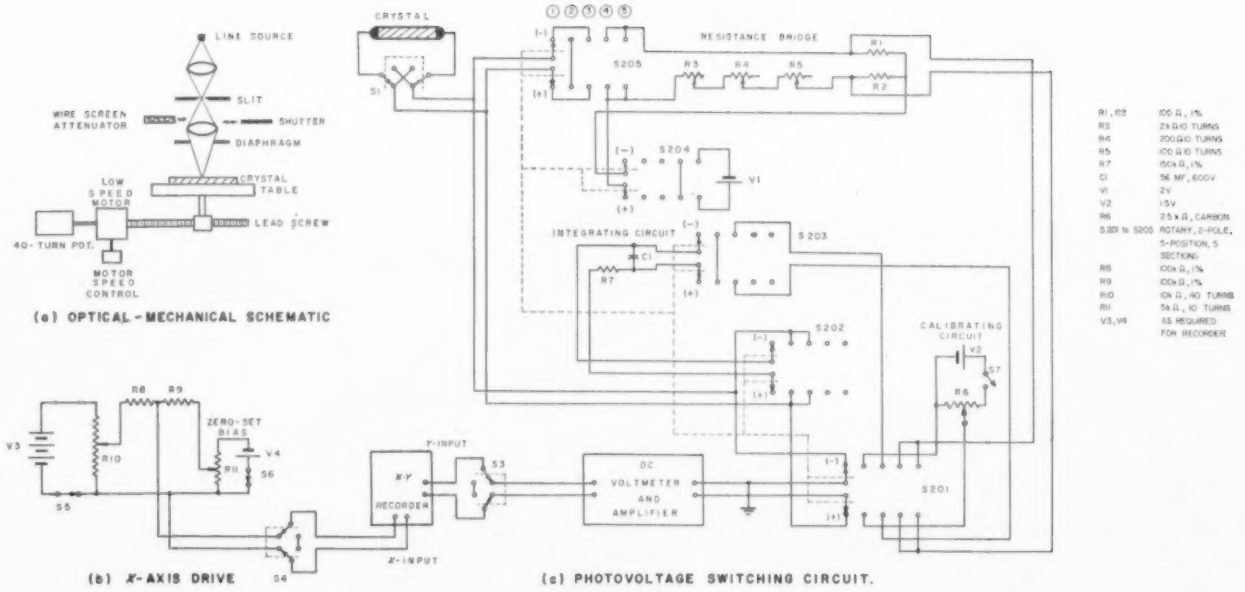


FIG. 2. Schematic of photovoltage scanning system.

position No. 4) any "short-circuit voltage" from the crystal, as is usually the case when the crystal is illuminated, then the bridge is balanced (in position No. 5) to the point where the voltmeter indicates that short-circuit voltage.

In order to enable a simple quantitative estimate of the resistivity along the filament by the procedure just described, the experimental conditions should be chosen so as to satisfy the assumptions underlying equation (7) or (8) as closely as possible. It is not easy however, to satisfy the condition that the resistivity is constant over a cross section. A two-point-probe resistivity measurement would, in this event, yield the average resistivity over the cross section. In order to obtain such an average from the photovoltaic scan, a line source is used so that a band of light extends across the width of the filament. Furthermore, the sample thickness is kept small compared with a diffusion length. If thicker samples are to be explored, penetrating light in the fundamental absorption edge should be used.

Ohmic contacts obtained by soldering are usually satisfactory. It is advisable to sand blast (or otherwise abrade) the surface of the crystal in the immediate vicinity of the contacts. This operation will reduce the filament effective lifetime in that region, and will in turn reduce the photovoltages due to contact barriers that may be present.

In practice the effective lifetime may not be uniform along the filament, giving rise to variations in ΔG from point to point. To correct this variation, ΔG is measured at two or three points along the filament and an average value is used in the subsequent calculation.

The sensitivity of the photovoltaic scan at a given light intensity is proportional to the filament lifetime. Therefore, filaments to be explored are always etched so as to increase the filament lifetime as much as possible. For a given filament lifetime the sensitivity increases with the light intensity. However, the latter should be limited by the requirement that $\Delta p \ll (p_0+n_0)$. This is easily insured by checking the linearity of the photovoltage with light intensity.

RESULTS

Typical results for rectangular filaments are shown in Figs. 3 and 4. In each figure, the upper and

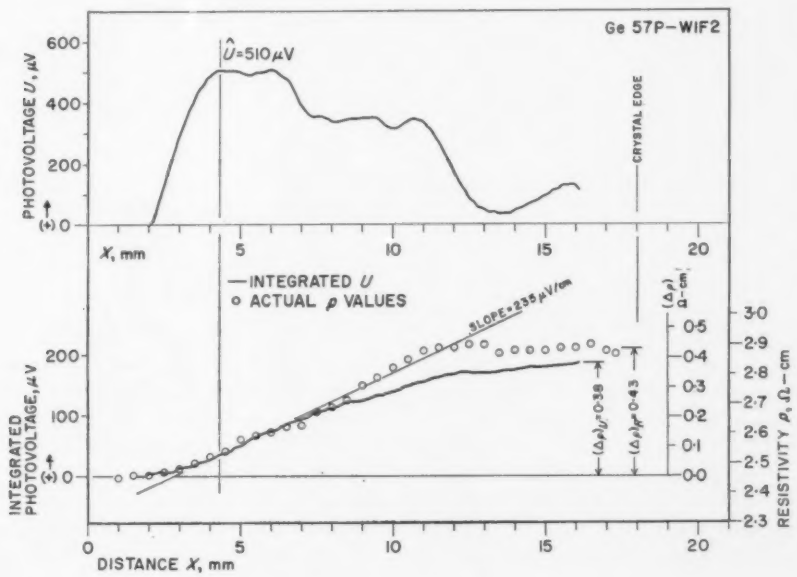


FIG. 3. Photovoltage and resistivity scan of a p-type germanium sample.

lower traces represent the direct and integrated photovoltage respectively, as recorded. The $(\Delta\rho)$ scale in Fig. 3 was derived from equation (8) using the measured ΔG and the dimensions of the filament. The ρ scale refers to the experimental points obtained from the two-point-probe measurement. It will be seen that the agreement is good, the error amounting to about 10 per cent of the maximum change in resistivity.

5 in the form of a three-dimensional isorho diagram. The wafer (p-type material) is in the form of a 16×20 mm rectangle. On the left-hand side, $(\Delta\rho)_U$ is derived from the photovoltaic scanning measurements. After the photovoltage measurements were completed, the rectangular wafer section was cut up into several rectangular filaments, and ρ vs. x was measured for each one using the two-point probe. From these latter measurements,

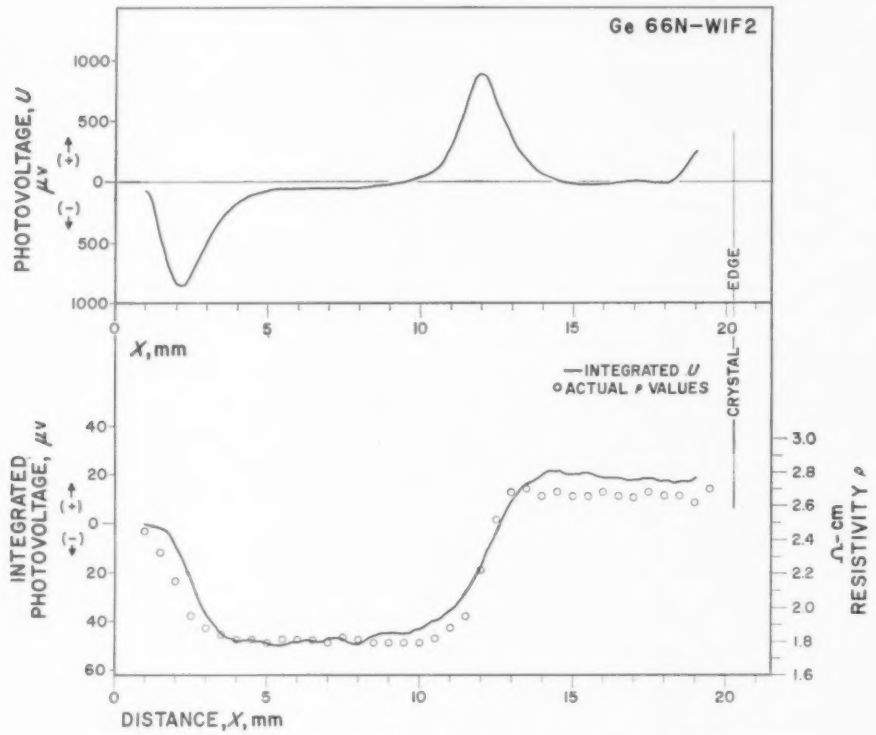


FIG. 4. Photovoltage and resistivity scan of an n-type germanium sample.

As mentioned earlier in this article, crystal wafers may also be explored by photovoltaic scanning. As many successive scans of the wafer are made (usually overlapping) as is necessary to cover the entire wafer. An isorho[1] map of the wafer is then prepared from the integrated photovoltage curves, and one then has a picture of the resistivity changes throughout the major portion of the wafer.

Typical results for a wafer are presented in Fig.

the $(\Delta\rho)_R$ surface was constructed and is shown in the right-hand side of Fig. 5. The heavy bars lying parallel to the y-axis of the right-hand surface indicate the width of each filament and the position which it occupied in the wafer. The striking similarity between both surfaces of Fig. 5 is self-evident. The overall error here amounts to about 10 per cent of $(\Delta\rho)_R$. Mere inspection of the $(\Delta\rho)_U$ surface discloses that the most uniform sample that may be obtained from this particular wafer

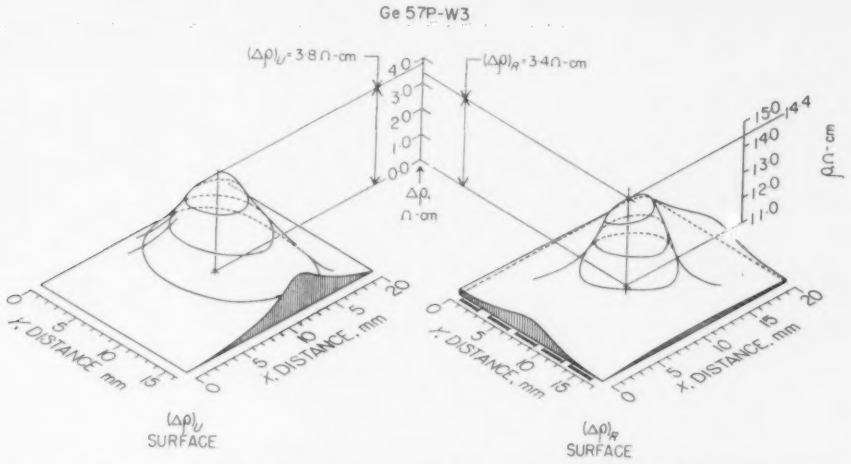


FIG. 5. Isorho surfaces for a *p*-type germanium wafer constructed from photovoltage scanning—$(\Delta\rho)_U$, and probe measurements—$(\Delta\rho)_R$.

will be from the region parallel to the *y*-axis and adjacent to $x = 0$. This is amply corroborated by the $(\Delta\rho)_R$ surface.

CONCLUSION

Photovoltaic scanning affords a fast, continuous and non-destructive means of exploring resistivity variations in both filaments and crystal wafers. In regard to wafers it is the only way of making a detailed estimate of the resistivity variations before cutting it into smaller samples. In spite of the fact that the results in this case are quite approximate, adequate information can be gleaned from the results to allow determination of those regions from which rectangular filaments of best uniformity may be obtained.

The accuracy as actually determined by measurements on a large number of samples lay between a few per cent and a few tens of per cent. The main source of error seemed to be in the non-uniformity of filament lifetime. In those filaments where the lifetime was found to be fairly uniform, the error was not higher than 3 per cent. Obviously, it is possible to correct for the non-uniformities in lifetime by taking a scan of ΔG as well as of U. However, such a procedure will appreciably increase the time of measurement and seems impractical.

Photo-scanning is feasible under conditions where ordinary probe methods would, at the very least, be quite difficult. Experiments where the sample is in vacuum or held at liquid-nitrogen temperature are two such instances.

Acknowledgements—The authors are indebted to Mr. ALFRED WILLIS and Mr. LOU LETENDRE of this laboratory for their assistance in preparing the samples.

REFERENCES

1. J. OROSHNIK and A. MANY, *J. Electrochem. Soc.* **106,** 360 (1959).
2. J. TAUC, *Czech. J. Phys.* **5,** 178 (1955).
3. W. SHOCKLEY, *Electrons and Holes in Semiconductors.* Van Nostrand, New York (1950).
4. W. SHOCKLEY, *Electrons and Holes in Semiconductors* p. 76. Van Nostrand, New York (1950).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 54–69. Printed in Great Britain

AVALANCHE INJECTION DIODES

A. F. GIBSON AND J. R. MORGAN

Royal Radar Establishment, Great Malvern, Worcs.

(*Received* 20 *June* 1959)

Abstract—A high-speed two-terminal negative-resistance device is described. A quantitative theory of the diode characteristics is developed and compared with experimental units.

Résumé—On décrit un élément à deux bornes et à résistance négative susceptible de fonctionner à très haute fréquence. Une théorie quantitative des caractéristiques de cette diode est décrite et comparée aux résultats expérimentaux.

Zusammenfassung—Es wird ein Zweielektroden-Bauelement mit negativem Widerstand und hoher Schaltgeschwindigkeit beschrieben. Eine quantitative Theorie der Dioden-Kennlinie wird entwickelt und mit experimentellen Messungen verglichen.

1. INTRODUCTION

GUNN[1] has described the avalanche injection effect in semiconductors and has shown that a two-terminal device showing negative-resistance properties can be made in germanium. Physically, this effect can be described most simply by considering a bar of n-type germanium with $n+$-(non-injecting) end connections through which is passed a steadily increasing current. The electric field E in the specimen rises steadily until a critical field (the avalanche field E_2) is reached at which electron–hole pair production occurs by collision ionisation, as in a gas discharge. In general the cross-section of the bar will not be uniform and the avalanche will occur at some arbitrary position along the bar. If the specimen is n-type, the holes produced in the avalanche drift towards the negative terminal and increase the conductivity of the bar near the negative end, thus reducing the field at this end. If the externally applied voltage remains approximately constant, the field at the positive end is enhanced (by an amount depending on the external circuit resistance) and the avalanche generation rate at the positive end will be increased. In the limit the field will be concentrated in a thin region near the positive electrode, the remainder of the bar being so heavily injected that only a small or negligible field exists in this region.

If the length of the bar is L_1, the voltage drop V_T at or just before the avalanche point will be

$$V_T \approx L_1 E_2 \tag{1}$$

where E_2 is the critical avalanche field. The rate of ionisation increases extremely rapidly with field for fields greater than E_2 so, after avalanching has occurred, the field in the avalanche region will still be about E_2. Hence the voltage drop across the avalanche region will be given approximately by

$$V_s \approx L_2 E_2 \tag{2}$$

where, as explained above, L_2 is less than L_1. If, therefore, the voltage drop across the remainder of the bar is small an extended negative resistance can be expected.

A further property of electrons in germanium which is important in the present connection is the saturation of the electron drift velocity at high electric fields first observed by RYDER[2]. When the electric field exceeds about 1 kV/cm the electron velocity (and hence the current) in n-type germanium becomes substantially independent of the electric field, up to the avalanche field (GUNN[3]). This effect facilitates the observation of a negative resistance due to avalanche injection in two ways, namely:

(i) the current density at the avalanche field, at least in a linear system of the type considered

above, is about 100 times less than that expected from Ohm's law.

(ii) The decrease in field per unit carrier injected from the avalanche is very considerably greater than that expected from Ohm's law, thus enhancing the positive feed-back process.

In practice a linear structure of the type described is very difficult to fabricate. As the value of E_2 is about 10^5 V/cm in germanium, a bar of length less than 10^{-3} cm (0·4 thou) must be used if V_T is not to be excessive [equation (1)]. If the current at or about the trigger voltage V_T is not to be unduly large, the cross-section must be of similar dimensions. Even so, the power dissipation per unit volume is considerable. In view of these considerations we have studied an alternative structure, namely a small, approximately hemispherical n_+-alloy "dot" on a relatively large crystal of n-type germanium or silicon (GIBSON and MORGAN[4]) with a view to assessing the practicability of making a two-terminal negative-resistance device using the avalanche injection effect. It is well known that a small hemispherical contact is the most favourable geometry from the point of view of heat dissipation and the small area ensures a high electric field at the contact for relatively modest currents. In addition this structure is relatively simple to reproduce, as the properties of only one junction are important.

In Sections 2 and 3 of this paper we describe the fabrication and general properties of "small dot" avalanche diodes. In Section 4 a theory of the trigger point is developed which indicates how the trigger voltage and current vary with crystal resistivity, alloy dot radius and temperature. The theory is compared with the behaviour of experimental diodes. The sustaining voltage V_S is discussed in Section 5, and in Section 6 the paper concludes with a description of a three-terminal device which is essentially an avalanche diode with an additional triggering electrode.

2. DIODE FABRICATION

Diodes have been made using n-type germanium and n- and p-type silicon ranging from twenty to several hundred Ω-cm resistivity. Early experience showed that the trigger voltage increased with increasing resistivity and that 40 Ω-cm material represented a reasonable compromise.

By far the largest number of diodes have been made using 20, 30 or 40 Ω-cm n-type germanium. Care is taken to use grown crystals having a low radial variation of resistivity, about ±2 Ω-cm. Pieces of germanium about 0·04 in. square by 0·015 in. thick are etched down to 0·012 in. thick in a fast CP4 type etch to remove the worked layer and polish the surface.

Using the temperature-reversal technique described by MASON and TAYLOR[5] two gold/0·75 per cent antimony-alloy wires, one 0·010 in. diameter and the other 0·002 in. diameter, are alloyed on to opposite sides of the germanium dice. Diodes have been made on dice cut parallel to ⟨111⟩ and ⟨100⟩ crystal planes, the latter proving more suitable for the final etching process as measured in terms of yield. Alloying on a ⟨100⟩ face gives a pyramid-shaped alloy front, bounded by ⟨111⟩ planes, which facilitates a reduction in contact area by etching after alloying.

The smaller wire is crimped after alloying and the unit mounted on a suitable header. Using a CP4 type etch incorporating the strongest available hydrofluoric acid, the mounted units are etched for 20 sec, washed in de-ionised water, dried in a stream of hot nitrogen, and their resistance measured at a low voltage. The low-voltage resistance is determined by the spreading resistance of the smaller n_+-alloy contact and if the alloy dot is approximately hemispherical is given by

$$R_0 = \frac{\rho}{2\pi r_0} \qquad (3)$$

where ρ is the resistivity of the material and r_0 the radius of the contact. If necessary, etching is continued until a desired value of r_0, and hence R_0, is achieved. Units are then dipped in a thick silicone varnish and encapsulated in a can containing a thick silicone grease.

3. DIODE CHARACTERISTICS

Typical characteristics of a germanium and a silicon diode fabricated as described are shown in Figs. 1 and 2 respectively. In this section we shall enumerate briefly the more important properties of such diodes. The trigger point is defined as the point at which the slope resistance dV/dI becomes zero, and subsequently negative, while the portion of the characteristic in which the current is greater than the trigger current is referred to as the sustaining region. Some diode properties, for example

the variation of trigger voltage with temperature or resistivity, are amenable to theoretical analysis. Such analysis and the comparison of theory and experiment form the subject matter of subsequent sections.

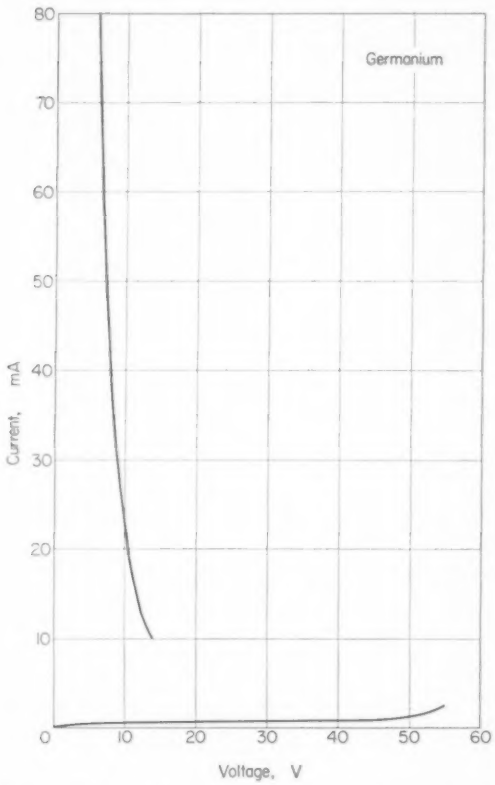


FIG. 1. Characteristic of a typical germanium avalanche diode.

(a) The trigger voltage V_T is found to increase slowly with crystal resistivity and dot radius r_0. The resistance at the trigger point is significantly greater than the resistance at low voltage, typically by a factor of two. This factor arises from a combination of the saturation of the drift velocity at high fields, together with a "majority carrier accumulation" effect (Section 4). In germanium units, the trigger voltage is found to be substantially independent of temperature up to about 140°C and then to fall slowly. For silicon diodes the critical temperature is, of course, considerably

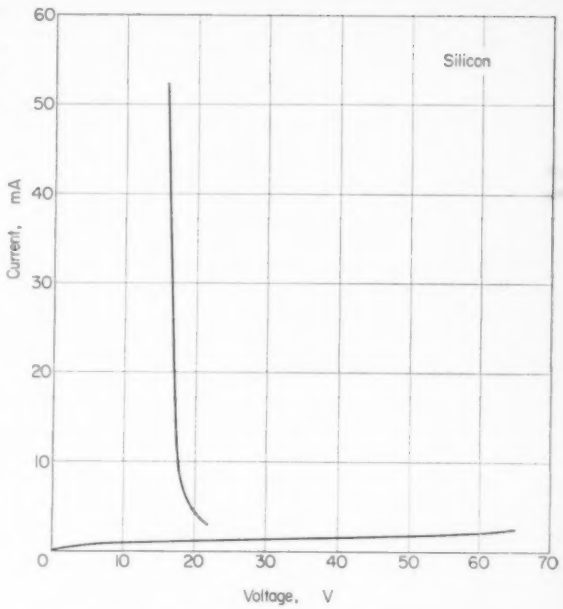


FIG. 2. Characteristic of a typical silicon avalanche diode. Compared with germanium, silicon units have a higher sustaining voltage, lower minimum sustaining current and less curvature near the trigger point.

higher. The theoretical interpretation of all these features is given in Section 4.

(b) The sustaining voltage at a given current is substantially independent of temperature, dot radius and crystal resistivity, provided the latter is not too low. The fraction of the sustaining current carried by holes can be estimated experimentally and is typically 20–30 per cent. The theory of the sustaining region is appreciably more difficult than that of the trigger voltage as the avalanche is transversely unstable, that is the avalanche does not occur uniformly over the alloy dot but is concentrated into a small region—a phenomenon familiar in gas discharges. These aspects are discussed more fully in Section 5.

(c) Considered as bi-stable devices, avalanche injection diodes can be switched "on" and "off" at relatively high speeds with low-energy pulses. Switching speed in bi-stable circuits can be discussed only in terms of the circuits used and some typical circuits are shown in Fig. 3.

Fig. 3(a) shows the simplest possible bi-stable circuit, together with a "load-line" characteristic

showing the two stable operating points for the circuit. The input and output condensers are kept small ($\sim 100\,\mu\mu$F) and a small resistor R_{L_2} ($\sim 100\,\Omega$) is included to limit the current surge through the diode on switching. R_{L_1} is typically a few kilo-ohms. This circuit suffers from a number of practical disadvantages, however, namely that the input and output connections are common, positive- and negative-going switching pulses (not generally available in computer circuits) are required for switching "on" and "off", and the resistor in the earth lead reduces the available output voltage.

Two of these disadvantages are overcome by the circuit shown in Fig. 3(b). In this circuit the diode V_1 is conducting all the time except when the "turn-on" pulse is applied. The turn-on pulse reverses the bias on V_1, so that a high-impedance input is presented to the turn-on pulse, until such time as the avalanche injection diode triggers. When triggering occurs the voltage drop across the avalanche diode decreases sufficiently to bias V_1

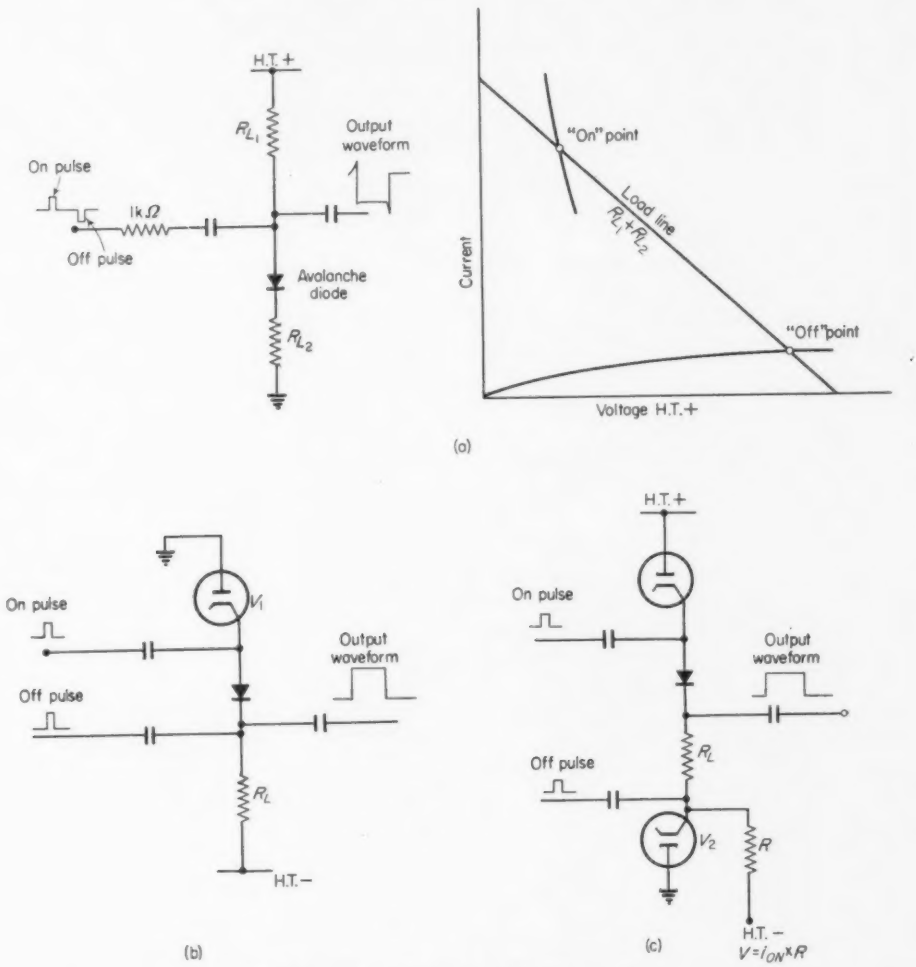


FIG. 3. Avalanche diode switching circuits.

forward again, even if the turn-on pulse is still being applied. The "turn-on" input is therefore very effectively isolated from the output circuit, "on", "off" and output pulses are of the same polarity and if the diode V_1 has a low forward resistance the decrease in output amplitude is negligible. Naturally a low hole-storage diode is required for V_1 and diodes of the VX3286 type have been found to be satisfactory.

At the cost of increased complexity, isolation of both inputs from the output can be obtained as indicated in Fig. 3(c). The diode V_2 plays a similar role to that of V_1 in as much as it is normally conducting and is closed by the "off" pulse until such time as the avalanche diode turns off. In this arrangement, very good isolation between input and output, together with a high input impedance, has been achieved. Both circuits 3(b) and 3(c) can, of course, be inverted if negative-going trigger pulses are desired. Alternatively both V_1 and V_2 of Fig. 3(c) may be joined in series and a single input point is used to which both positive (on) and negative (off) pulses may be applied.

The switching characteristics of these circuits have been examined experimentally. The following general features have been noted.

(i) The time required to turn the circuit "on" is substantially independent of the amplitude of the "on" pulse or the bias conditions and is typically 10–20 mμsec, as measured on a "Tektronix" type 545 Oscilloscope. This time constant appears to be determined primarily by the circuit used and as the specified rise time of this Oscilloscope is 12 mμsec the avalanche diodes can probably be switched on in less than 10 mμsec. In practical circuits, however, the switching time is largely determined by the time taken to charge stray capacities associated with succeeding stages.

(ii) The time required to turn the circuit "off" decreases with increasing trigger-pulse amplitude and increases slowly with the magnitude of the "on" current of the avalanche diode. A typical turn-off time is 30–40 mμsec.

(iii) The amplitude of the pulses required to turn the diodes "off" and "on" increases and decreases respectively as the HT voltage is increased. It is convenient to operate under conditions such that the "on" and "off" trigger pulses required are of equal amplitude and about 5 V.

4. THE TRIGGER POINT

In switching circuits, the most important parameter of any two-terminal negative-resistance device is the trigger voltage.

(a) *The trigger point at room temperature. Theoretical*

The pre-trigger or "off" state of an avalanche diode may be specified by three quantities, namely the resistance at low voltages R_0, the trigger voltage V_T and the trigger current I_T. We shall now deduce theoretical expressions between these quantities subject to the following assumptions.

(i) That the material is n-type germanium and that the current due to thermally generated holes can be ignored. This assumption cannot be justified at high temperatures, as discussed in Section 4(c).

(ii) That the electron drift velocity is proportional to the electric field up to $E = E_1$ but independent of field in the range $E_1 < E < E_2$. The degree of approximation represented by this assumption may be assessed from GUNN's results[3] which indicate that the drift velocity of electrons in germanium increases by a factor of about 2·5 when the field is increased 100 times.

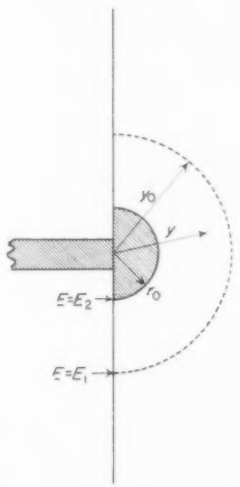


FIG. 4. Model for theory of avalanche diodes.

(iii) That the small alloy dot is hemispherical. This is probably the worst approximation used. With pyramid-shaped alloy junctions, for example, it is found experimentally that the trigger

voltage decreases if the pyramid point is made sharper. However, three parameters are required to specify any contact geometry other than hemispherical and as only three diode parameters are measured agreement between experiment and a geometrically more detailed model would not be significant.

Referring to Fig. 4, r_0 is the radius of the hemispherical alloy dot and y_0 is the radius at which the electric field equals E_1. Let y be any radius. Clearly the current density and electric field increase with decreasing radius, the maximum value occurring at $y = r_0$. To determine the trigger conditions we put the field at r_0 equal to the avalanche field E_2. For $r_0 < y < y_0$ the electrons drift at a constant velocity v_s independent of field. For $y > y_0$ the electron drift velocity is proportional to the electric field and Ohm's law applies.

The trigger voltage is then given by

$$V_T = \int_{r_0}^{y_0} E(y)\,dy + \int_{y_0}^{\infty} E(y)\,dy = V_1 + V_2 \tag{4}$$

where the second, "spreading resistance" term is given by

$$V_2 = \frac{I_T \cdot \rho}{2\pi y_0} \tag{5}$$

where I_T is the trigger current and ρ is the resistivity of the material.

In the volume defined by $r_0 < y < y_0$, however, there is a significant space charge. This arises from the saturation of the electron drift velocity. The current density must increase as the radius y decreases but the electron drift velocity cannot exceed v_s so the electron density n must increase to carry the current. The variation of n with y is given by

$$I = 2\pi y^2 nqv_s + 2\pi y^2 \cdot qD\left(\frac{dn}{dy}\right) = 2\pi y_0^2 Nqv_s \tag{6}$$

where N is the density of donor impurities in the bulk material, assumed equal to the electron density in the absence of space charge ($y \geqslant y_0$), q is the electronic charge and D the electron diffusion constant.

If the diffusion term is small, n is given by

$$n = N\left(\frac{y_0}{y}\right)^2 \times \times\left[1+2\left\{\frac{1}{\left(\frac{yv_s}{D}-1\right)} + \frac{1}{\left(\frac{yv_s}{D}-1\right)^2} \ldots\right\}\right] \tag{7}$$

Clearly the diffusion term is most important when y is small, that is equal to r_0. We may neglect diffusion entirely if

$$D \ll \frac{r_0 v_s}{3} \tag{8}$$

The value of v_s is about 6×10^6 cm/sec (GIBSON and GRANVILLE[6], GUNN[3]) and in practice r_0 is not less than 10^{-3} cm. The value of D for electrons at high electric fields is not known but is unlikely to be as large as 3000 cm^2/sec (see Section 5) so, in order to proceed, we shall assume that condition (8) is satisfied.

Hence

$$n = N(y_0/y)^2$$

As the system under consideration has spherical symmetry, Poisson's equation (in rationalised units) may be written in the form

$$\frac{1}{y^2}\cdot\frac{d(y^2E)}{dy} = -\frac{q}{\kappa}(n-N) = -\frac{Nq}{\kappa}[(y_0/y)^2-1] \tag{9}$$

where κ is the dielectric constant of the material.

Integrating equation (9) with the boundary condition that $E = E_1$ at $y = y_0$ gives

$$y^2E - y_0^2E_1 = \frac{1}{3\rho\mu\kappa}(y_0-y)(2y_0^2 - yy_0 - y^2) \tag{10}$$

where $1/\rho\mu$ has been written for Nq, μ being the low field electron mobility.

V_1, equation (4), is given by

$$V_1 = \int_{r_0}^{y_0} E(y)\,dy$$

Integration with the boundary condition $E = E_2$

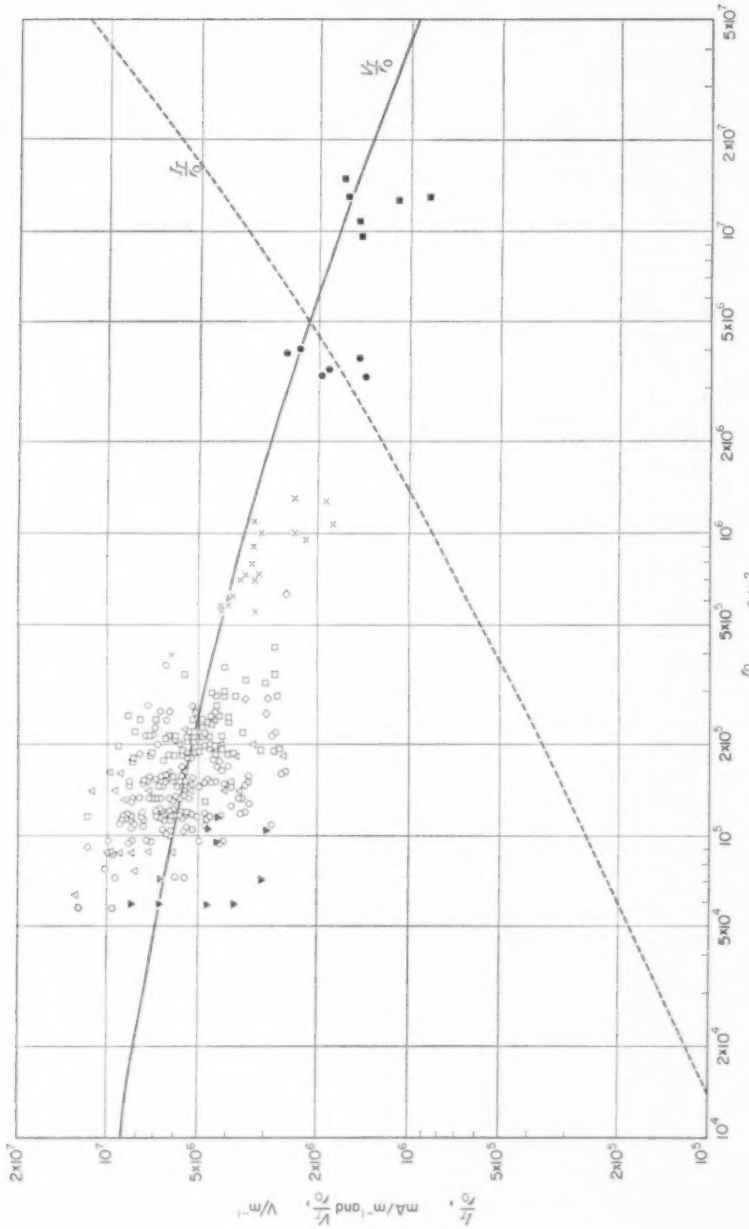


FIG. 5(a). Theoretical curves of V_T/r_0 and I_T/r_0 as functions of $r_0/3\rho\mu\kappa$. As r_0 is deduced from I_T all diodes must fall on this curve. Crystal resistivities in Ω-cm:

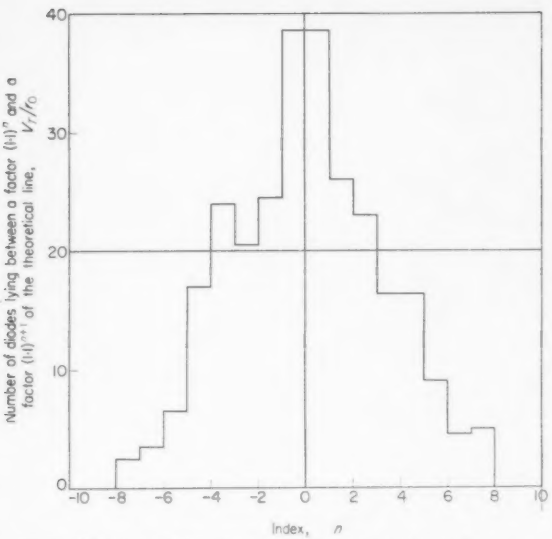

Number of diodes lying between a factor (1·1)^n and a factor (1·1)$^{n+1}$ of the theoretical line, V_T/r_0
40
30
20
10
0
-10 -8 -6 -4 -2 0 2 4 6 8 10
Index, n

FIG. 5(b). The histogram indicates deviation of experimental diodes from theoretical line for V_T/r_0.

at $y = r_0$ (the condition for an avalanche) gives

$$V_1 = E_1\left(\frac{y_0^2}{r_0} - y_0\right) + \\ + \frac{1}{3\rho\mu\kappa}[2y_0^3/r_0 - 2y_0^2 - 3y_0^2 \ln(y_0/r_0) + \tfrac{1}{2}(y_0^2 - r_0^2)] \quad (11)$$

where y_0 is given by

$$r_0^2 E_2 - y_0^2 E_1 = \frac{1}{3\rho\mu\kappa}(y_0 - r_0)^2(2y_0 + r_0) \quad (12)$$

Finally the trigger current I_T is given by equation (6) as

$$I_T = \frac{2\pi v_s y_0^2}{\rho\mu} \quad (13)$$

where y_0 is given by equation (12) as before.

(b) *The trigger point at room temperature. Experimental*

The above equations for V_T and I_T do not have simple analytical solutions and have been solved numerically.* For graphical representation it is convenient to determine the quantities (V_T/r_0) and (I_T/r_0) as functions of $r_0/3\rho\mu\kappa$; the equations may then be represented by a single curve. In addition we have equation (3) relating to the diode resistance at low voltages, R_0, and the alloy dot radius, r_0. If we assume that κ, ρ, μ, v_s, E_1 and E_2 are known constants of the material the only unknown diode parameter is r_0 and any of the three measured quantities (V_T, I_T or R_0) may be used to determine r_0 for each diode. We have chosen to determine r_0 from the trigger current I_T and use the remaining two quantities to compare theory and experiment.

The theoretical curves shown in Figs. 5(a) and 6 assume the following numerical values of the constants:

$$\kappa = 16 = 1{\cdot}4 \times 10^{-10} \text{ F/m}$$

$$E_1 = 10^3 \text{ V/cm}$$

$$E_2 = 10^5 \text{ V/cm}$$

$$v_s = 6 \times 10^6 \text{ cm/sec}$$

$$\mu = 3800 \text{ cm}^2/\text{V per sec}$$

The values of E_1 and E_2 are based on measurements by GUNN[3] and others. In practice the results are relatively insensitive to the choice of E_1. The value of v_s is a mean of results by RYDER, GUNN and others, tabulated by GIBSON and GRANVILLE[6]. The value of v_s appears linearly in the expression for trigger current but is relatively unimportant in the trigger voltage.

There are naturally no experimental points on the theoretical curve for I_T/r_0 as a function of $r_0/3\rho\mu\kappa$ given in Fig. 5(a) as all diodes must lie on this curve by the definition of r_0. Each experimental point on the V_T/r_0 curve represents a diode and all recorded diodes have been included. The spread of experimental values is rather wide, but the general agreement between theory and experiment, over a range of nearly a factor of 1000 in $r_0/3\rho\mu\kappa$, is as good as can be expected. That the theoretical line is also the best mean line is shown by the histogram in Fig. 5(b): about 30 per cent of all diodes lie within ± 10 per cent of the theoretical line and about 90 per cent lie within ± 60 per cent. As the theoretical value of V_T/r_0 for a given $r_0/3\rho\mu\kappa$ is fairly sensitive to the choice of E_2 this result represents an accurate determination of the effective avalanche field for germanium.

It has been noticed that, for any given resistivity,

* We are indebted to Mr. P. M. WOODWARD and Miss R. HENSMAN for their assistance in these computations.

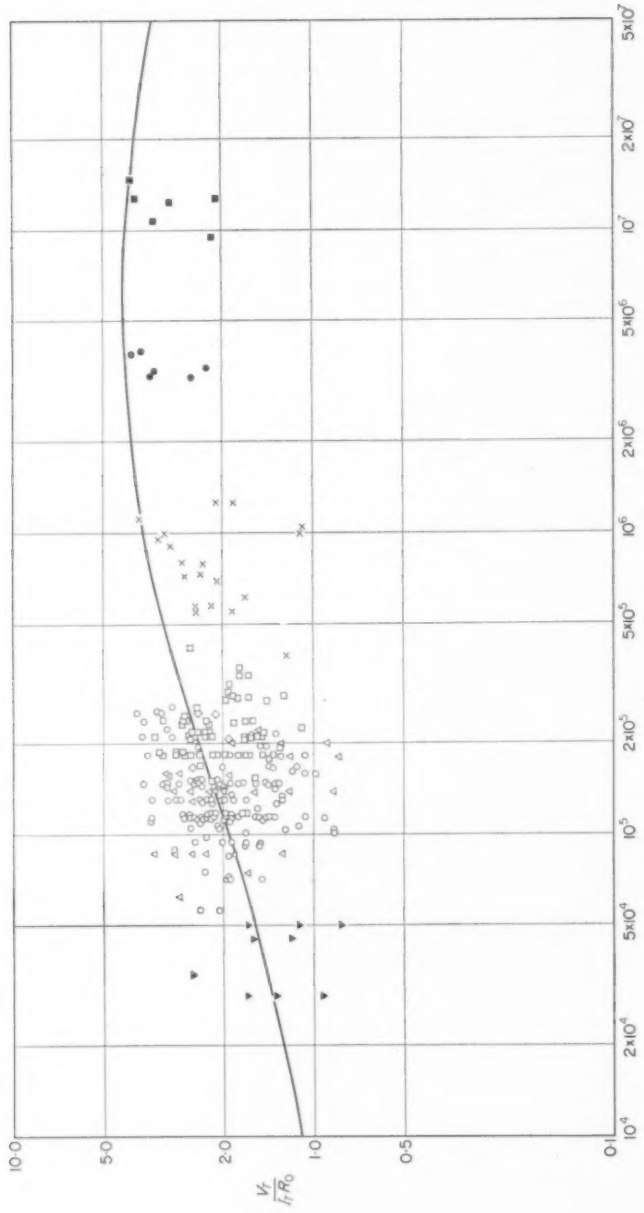


FIG. 6. Theoretical curve of $V_T/I_T R_0$ as a function of $r_0/3\rho\mu\kappa$. Same experimental diodes as shown in Fig. 5.

the experimental diodes lie on a line of somewhat steeper slope than the theoretical line and an empirical curve may be drawn through the points such that the spread from this line is much less than that indicated by Fig. 5(a). Hence at least part of the spread around the theoretical V_T/r_0 curve in Fig. 5(a) is not random but due to the inadequacy of the theory to give an accurate description of the variation of V_T/r_0 with r_0. Presumably this arises from the assumption of a simple geometry. Thus the variation of V_T with r_0, predicted by theory to be slower than linear, is in fact very slow. This is at least fortunate from the practical point of view as it increases the reproducibility of V_T from unit to unit and in practice V_T is certainly the most important diode parameter.

In Fig. 6 we show the quantity V_T/I_TR_0 as a function of $r_0/3\rho\mu\kappa$. This quantity measures the ratio of the resistance at the trigger point to that at low voltage and is a measure of the curvature due to the saturation of the electron drift velocity. In a linear structure V_T/I_TR_0 would be about 100 (Section 1) but is much lower, both theoretically and experimentally, in a radial structure owing to the accumulation of electrons near the contact, which effectively reduces the resistivity at high voltages. The experimental points in Fig. 6 are, of course, the same diodes as those of Fig. 5. Again agreement between theory and experiment is adequate, though the experimental points appear to be generally below the theoretical curve. It is believed that this is due to a small amount of avalanche generation of holes at voltages below the trigger voltage. Certainly the current starts to increase rapidly with voltage very near the trigger point even though the slope resistance remains positive (see Fig. 1). This feature is not included in the theoretical treatment.

Two final points should be noted, namely:

(i) that the theoretical equations given above are still valid for near intrinsic material, provided that the resistivity is interpreted as $1/Nq\mu$, where N is the number of impurities. Some of the diodes shown in Figs. 5 and 6 were made on material having an effective resistivity of 80 Ω-cm.

(ii) n- or p-type silicon diodes have very similar trigger characteristics (V_T and I_T) to germanium units of comparable resistivity and dot radius. The values of the electron or hole drift velocities in silicon are not so well known as those in germanium, so no quantitative analysis has been attempted. However the availability of very high resistivity silicon* (~ 5000 Ω-cm) has allowed the fabrication of diodes with trigger voltages in excess of 200 V, with quite modest trigger currents.

(c) *The trigger point at elevated temperatures. Theoretical*

At high temperatures the generation rate and density of holes in the space-charge region can no longer be neglected in a calculation of the trigger voltage. As the dot is positive on n-type material the generated holes are swept away from the dot at their saturated drift velocity and we may estimate the hole density $p(y)$ by equating the rate of generation between r_0 and y to the rate at which holes leave an element dy thick at radius y.

The rate of generation of holes per unit volume in n-type material is given approximately by $n_i^2/N\cdot\tau$, where n_i is the intrinsic electron or hole density and τ is the carrier lifetime. n_i^2 varies exponentially with temperature and is the dominant temperature-dependent term.

The generation rate between y and r_0 is then

$$\frac{n_i^2}{N\tau}\left(\frac{4}{3}\pi\right)(y^3-r_0^3)$$

which may be equated to the number leaving an element at radius y per unit time, given by

$$\frac{p\cdot 2\pi y^2\cdot dy}{dy/v_s}=2\pi y^2 v_s\cdot p.$$

Therefore

$$\begin{aligned} p(y) &= \frac{2n_i^2}{3\tau\cdot N\cdot v_s}[y-r_0^3/y^2] \\ &= Q(y-r_0^3/y^2) \end{aligned} \tag{14}$$

where Q has been written for brevity.

If the saturated drift velocity v_s is the same for electrons and holes (RYDER[2]) and the diffusion term is neglected as before [Section 4(a)], the current I at y and at y_0 is now given by

$$I=2\pi y^2\cdot qv_s(n+p)=2\pi y_0^2\cdot q\cdot v_s(n_0+p_0) \tag{15}$$

* We wish to thank Standard Telecommunications Laboratories, Enfield for supplying this material.

equation (15) being equivalent to equation (6) of Section 4(a).

At $y = y_0$ there is no space charge so, by electrical neutrality,

$$n_0 = p_0 + N \tag{16}$$

where p_0 is given by equation (14) with $y = y_0$.

Equations (15) and (16) give the variation of n with y, as in section 4(a). Poisson's equation now becomes

$$\frac{1}{y^2} \cdot \frac{d(y^2E)}{dy} = \frac{-q}{\kappa}[n-p-N] \tag{17}$$

Integrating twice, with the same boundary conditions as before [Section 4(a)], gives the voltage

$$= \frac{q}{\kappa}[N(r^3/3 - y_0^2 r + \tfrac{2}{3} \cdot y_0^3) + 2Q(r^4/4 - y_0^3 r + \tfrac{3}{4} y_0^4)] \tag{19}$$

The trigger current I_T is now given by

$$I_T = 2\pi q v_s [y_0^2 N + 2Q(y_0^3 - r_0^3)] \tag{20}$$

Equations (18), (19) and (20) reduce to (11), (12) and (13) respectively when $Q = 0$, corresponding to a low temperature at which n_i^2 is very small.

The numerical solution of equations (17) and (18) is tedious, and has only been attempted for $\rho = 40$ Ω-cm and one value of r_0, corresponding to $r_0/3\rho\mu\kappa = 1{\cdot}6 \times 10^5$ C/m² and a trigger voltage at low temperatures of about 47 V. In addition we have assumed that τ, μ, v_s, E_1 and E_2 are invariant

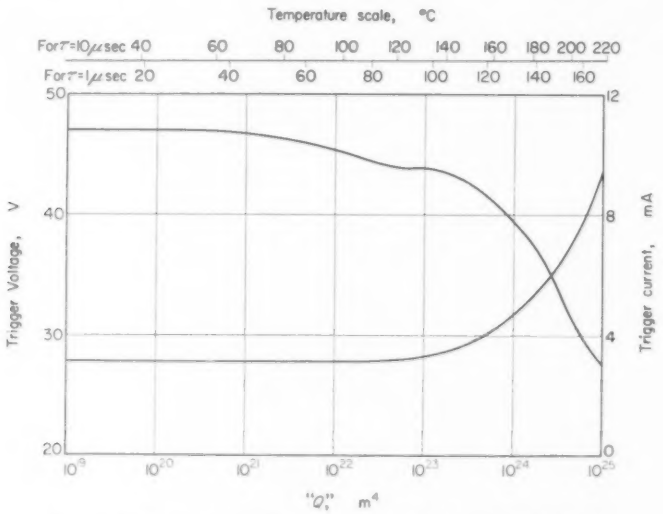


FIG. 7. Theoretical curves of V_T and I_T as functions of temperature for a germanium avalanche diode made with 40 Ω-cm material.

drop across the space-charge region as

$$V = E_1(y_0^2/r_0 - y_0) + \frac{qN}{3\kappa}[2(y_0^3/r - y_0^2) - 3y_0^2 \ln(y_0/r) + \tfrac{1}{2}(y_0^2 - r_0^2)] + \frac{Qq}{\kappa}[\tfrac{3}{2}(y_0^4/r - y_0^3) - 2y_0^3 \ln(y_0/r) + \tfrac{1}{6}(y_0^3 - r^3)] \tag{18}$$

where y_0 is given by

$$r_0^2 E_2 - y_0^2 E_1$$

with temperature, only Q being temperature dependent. This approximation allows V to be deduced as a function of Q without any *a priori* knowledge of the lifetime τ. Clearly the relevant value of τ is the lifetime of the material within a radius 2 or 3 r_0 of the dot, and this may be markedly less than the initial lifetime of the crystal before alloying.

The result of this computation* is shown in

* We are indebted to Mr. J. B. ARTHUR for computing this curve.

Fig. 7, in which two temperature scales, corresponding to $\tau = 1$ and 10 μsec respectively, have been included. It will be seen that the temperature at which a significant fall in V_T occurs is relatively insensitive to the value of τ. The inflexion in the curve relating V_T with Q at about $Q = 5 \times 10^{22}$ is believed to be real and not a feature of computational or other approximations. No comparable inflexion occurs in curve of trigger current with temperature.

is about 10 μsec. The lifetime of the starting material is several hundred μsec so this value does not seem unreasonable.

From the practical point of view the slow variation in V_T or I_T with temperature is extremely important. If a 10 per cent decrease in V_T may be tolerated, operation up to 130–140°C ambient is permissible. Alternatively a very high power dissipation can be achieved at moderate ambient temperatures by the use of an efficient heat sink,

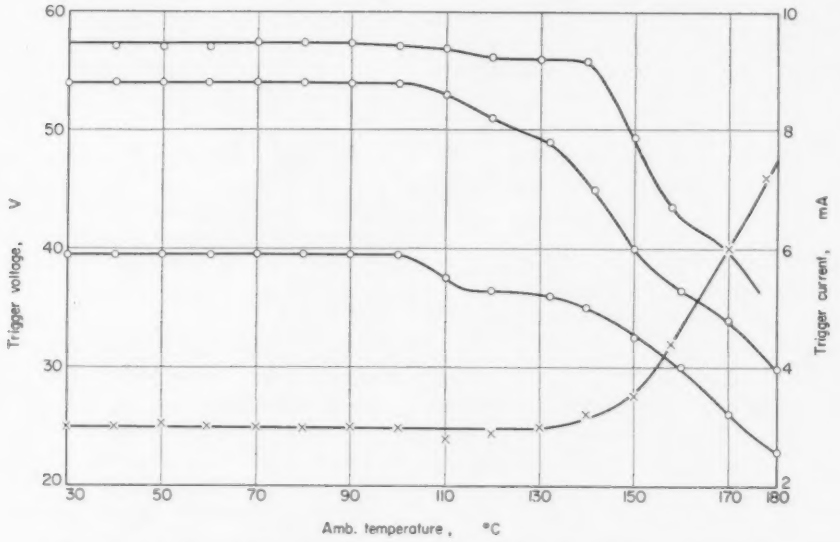


FIG. 8. Experimental curves of V_T and I_T as functions of temperature for typical germanium avalanche diodes made with 40 Ω-cm material.

(d) *The trigger point at elevated temperatures. Experimental*

Measurements of the variation of trigger voltage and trigger current with temperature have been made using a number of 40 Ω-cm germanium diodes and some typical results are shown in Fig. 8. The false origin of the trigger-voltage scale is the same in Figs. 7 and 8 so they are directly comparable. It will be seen that the agreement between theory and experiment is good and in particular there is experimental evidence of the inflexion in the V_T curve around 120°C. Comparison of the theoretical and experimental curves suggest that in the immediate neighbourhood of the alloy dot τ

the junction temperature being permitted to rise to 130–140°C. Considerably higher temperature ratings can, of course, be achieved by using silicon, in which n_i^2 is about 10^6 times smaller than in germanium.

5. THE SUSTAINING REGION

(a) *General characteristics*

For diode currents greater than the trigger current the slope resistance (dV/dI) is negative up to at least 100 mA and generally higher. At low currents (dV/dI) is typically $-10\,\text{k}\Omega$ and the characteristic is obscured by relaxation oscillations. The sustaining current at which $|(dV/dI)|$ falls

sufficiently for oscillations to cease, is a function of the test equipment, but is typically $2I_T$. At high currents (greater than 200 mA) the slope resistance is small but generally still negative, though a positive parasitic resistance, which can usually be ascribed to the base contact, is sometimes important. The extended range of the negative slope resistance of the diode can be useful in some switching circuits (see, for example, TOWNSEND[7]).

The value of the sustaining voltage at any specified current is found to be substantially independent of the alloy dot radius. For 95 per cent of diodes made on 20, 30 or 40 Ω-cm germanium the sustaining voltage at 25 mA lies between 8 and 11 V. The value of the sustaining voltage falls slightly at low resistivities and for diodes made with 1 Ω-cm material the average sustaining voltage is between 7·5 and 8 V. The sustaining voltage is thus inherently more reproducible than the trigger voltage.

(b) *Theoretical discussion*

A theoretical analysis of an injecting avalanche, which can be applied to the present problem, has been given by GUNN[8].

His assumptions were as follows:

(i) Planar geometry. The thickness of the avalanche region [L_2 of equation (2)] must be of the order of $1/\alpha$, where α is the avalanche ionisation constant (MCKAY,[9] MILLER[10]) and hence of the order of 10^{-5} cm. This is much less than the radius of any practical alloy dot, so planar geometry may be assumed to apply in the present case.

(ii) Carrier space charge in the avalanche region much greater than the impurity density, i.e. $I/_{\text{Area}}.v_s \gg N$. This assumption is justified in our case.

(iii) $\alpha_p = \alpha_n = \alpha$ and $v_{s_p} = v_{s_n} = v_s$.

(iv) The variation of α with field to be of the form

$$\alpha = \alpha_0 \exp(\lambda E) \qquad (21)$$

(v) Lateral stability of the avalanche, i.e. a uniform current density across the surface of the contact.

With these assumptions GUNN obtained, as an approximate solution,

$$V_s = 0{\cdot}414\left(\frac{4\pi\kappa v_s}{\lambda^3\alpha_0}\right)^{\frac{1}{3}} \cdot J^{-\frac{1}{3}} \qquad (22)$$

where λ and α_0 are defined by equation (21) and J is the current density. Hence, at a given current, V_s is proportional to the alloy dot radius r_0. This is contrary to the experimental result stated above. The crystal resistivity does not appear in equation (22) because of assumption (ii) above but it is clear that, on GUNN's model, V_s would decrease slightly in low-resistivity material, as observed.

That the sustaining voltage is approximately inversely proportional to the square root of the current is confirmed experimentally. All diodes examined can be represented, at least in the range 10–100 mA, by a relation of the form

$$V_s = \text{const.}I^{-n}$$

where n lies between 0·4 and 0·6, depending on the diode.

Numerical values of λ and α_0 can be estimated by fitting equation (21) to the experimental data on ionisation rate given by MILLER[10]. The absolute magnitude of the sustaining voltage at 25 mA may then be calculated from equation (22) assuming a typical alloy dot radius of 10^{-3} cm and avalanching to take place uniformly across the contact surface. This procedure indicates a sustaining voltage of about 70 V, which is an order of magnitude too large.

This discrepancy, together with the invariance of V_s with r_0, can be explained if it is assumed that, as in a gas discharge, the avalanche is laterally unstable and covers only a small fraction of the area of the contact. To obtain a sustaining voltage of about 10 V at 25 mA requires an effective avalanche area of radius about 10^{-4} cm and unless the alloy dot is smaller than this V_s will be independent of r_0.

(c) *The effective area of an injecting avalanche*

That the avalanche will in fact shrink to a small area may be shown by consideration of the equipotential lines around the contact. If the degree of avalanche injection is not uniform across the surface, the enhanced conductivity due to injection from a highly injecting region will so distort the equipotential lines as to increase the field, and hence the injection, in this region. Hence the stable configuration is a restricted area of avalanche. It may be anticipated that, by analogy with a gas discharge, the effective area of the avalanche will increase with current.

The effective avalanche area cannot shrink indefinitely but will be limited either by diffusion or space-charge considerations, whichever gives the greater radius. Space-charge effects will be important if the diameter of the avalanche region is about the same as its thickness in the field direction, i.e. 10^{-5} cm, but such a small area would lead to a much smaller sustaining voltage than observed. Hence we conclude that the area of the avalanche is diffusion-limited. If so, the radius of the avalanche region is simply given by equating the inward current due to the field (qnv_s) to the outward flow due to diffusion [$qD(dn/dr)$]. Hence $n = n_0 \exp(-v_s r/D)$ and we may define the effective radius of the avalanche region as D/v_s. However the diffusion constant of carriers in germanium in the presence of a high electric field has not been estimated, so far as we are aware, either experimentally or theoretically.

It would appear that a study of the sustaining region of an avalanche diode represents a method of studying this quantity so we shall now give a rough estimate of the diffusion constant at high electric fields.

(d) *The diffusion constant of hot electrons in germanium*

At high electric fields, electrons and holes in germanium gain energy from the electric field more rapidly than they can lose it to the lattice. Hence their average energy, or effective temperature, rises (SHOCKLEY[11]). We shall assume that the carriers have sufficient collisions between themselves to ensure that the energy distribution of the carriers is Maxwellian at all fields: at low fields this is obviously a good approximation, and in an avalanche region the carrier density is very high indeed, so that carrier collisions are very frequent. Hence we may use the Einstein relation

$$D_f = \mu k T_e / q$$

where D_f is the effective diffusion constant, k Boltzmann's constant and T_e the electron or carrier temperature. The mobility μ is determined by the rate of loss of momentum of the carriers and

$$\mu \approx q\tau/m$$

where τ is the momentum relaxation time and m the carrier effective mass.

At low electric fields, acoustic mode scattering dominates the rate of momentum loss. Hence:

$$\tau = l/\bar{v} \propto l/\sqrt{(T_e)}$$

where l is the mean free path and $\bar{v}$ is the average thermal velocity.

Hence

$$D_f = \frac{\mu k T_e}{q} = D_0 \cdot \mu_0 \frac{E}{v} \qquad (23)$$

where D_0 and μ_0 are the diffusion constant and mobility at zero field and v is the drift velocity in a field E.

At high electric fields, however, optical mode or intervalley scattering are dominant and τ is of the form

$$\tau \propto \frac{1 - \exp\left(-\frac{\hbar\omega}{kT}\right)}{\left(\frac{\Delta\epsilon}{kT} - 1\right)^{\frac{1}{2}}}$$

(HERRING[12]), but at high fields $\Delta\epsilon \gg kT$ and hence τ is still approximately given by

$$\tau \propto l/\sqrt{(T_e)}$$

where l is the mean free path under high-field conditions. Analysis of the experimental data on the variation of the ionisation constant α by MILLER[10] indicates that l is of about the same magnitude under avalanche conditions as at low fields. Hence equation (23) is still a fair approximation at the very highest fields. We shall therefore assume it to be valid at all fields.

The radius of the avalanche region is then given by

$$r = \frac{D_f}{v_s} \approx \frac{D_0 \mu_0 E}{v_s^2} \qquad (24)$$

$$\approx 10^4 \text{ cm when } E = E_2 = 10^5 \text{ V/cm}$$

and is clearly of the required order of magnitude to explain the observed value of the sustaining voltage.

As the area of the contact covered by the avalanche region is now a function of E (and hence current, as in a gas discharge) the integrations upon which equation (22) are based are strictly no longer valid. If the effective radius from equation (24) is included no simple analytical solution can

be obtained when α is defined by equation (21). As good a fit to MILLER's experimental data for α can be obtained however by writing

$$\alpha = \alpha_1 E^m$$

where $m = 6$ for germanium and $\alpha_1 = 1{\cdot}6 \times 10^{-38}$ $m^6\ V^{-6}$ from MILLER's data. Then, including an effective area defined by equation (24) we obtain, following GUNN's method,

$$V_s = \left(\frac{2}{m-5}\right)\left[\frac{I \cdot v_s{}^3}{\pi\kappa D_0{}^2\mu_0{}^2}\right]^{(1-m)/(m+3)} \times$$

$$\times \left(\frac{m+1}{4\alpha_1}\right)^{4/(3+m)} \text{(if } m > 5) \qquad (25)$$

≈ 7 V at 25 mA, which is of the required magnitude. It will be noticed that V_s is now proportional to (current)$^{-5/9}$ which is experimentally indistinguishable from the $-\frac{1}{2}$ law predicted by GUNN. GUNN's analysis indicates that the width of the avalanche region in the field direction decreases and the field increases slowly with increasing current. Similar results are obtained when the effective area of the avalanche region is assumed to be determined by equation (24), putting $m = 6$, namely:

The width of the avalanche	$\propto I^{-2/3}$
The electric field	$\propto I^{1/9}$
The effective area of the avalanche	$\propto I^{2/9}$

6. AVALANCHE TRIODES

By adding a p_+-alloy junction to an otherwise conventional n-type avalanche diode structure a number of new circuit arrangements become possible, which we have examined. In general any three-terminal minority carrier device will be limited in speed by carrier transit times between any pair of electrodes, and the present structure is no exception. Thus one virtue of the diode structure—high speed coupled with simplicity of fabrication—is lost. However, the development of modern transistor technology has largely overcome the difficulties of close junction spacings and the addition of a p_+-junction can considerably increase the versatility of the device so that some sacrifice in speed and increased cost may be justified.

To illustrate the properties of a p_+–n–n_+ structure (of which the n_+-junction is of small area) the following arrangements may be quoted as examples:

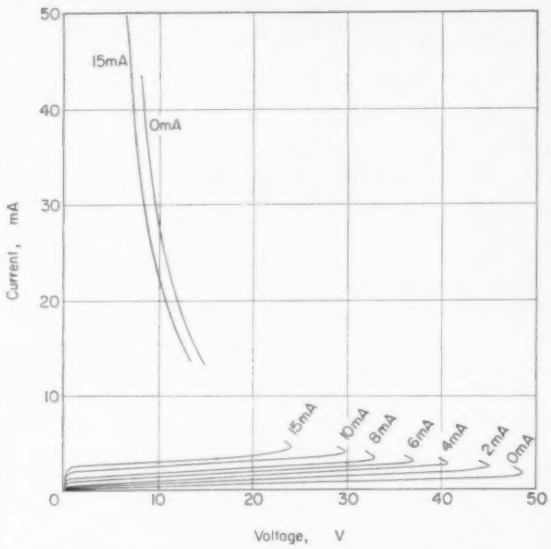


FIG. 9. "Thyratron" characteristic. The effect of minority carrier injection from a p–n junction on the trigger characteristics of an avalanche diode, (p–n junction currents indicated on diagram).

(i) *"Thyratron" type characteristic.* The n_+-avalanche dot is biased positive in the usual avalanche direction and the p_+-junction is also biased positive. Injection of holes from the p_+-junction increases the conductivity of the base region and hence decreases the trigger voltage of the avalanche diode. A typical static characteristic is shown in Fig. 9, each curve being an avalanche diode characteristic at a given value of p_+-junction current. Clearly positive pulses applied to the p_+-electrode will turn on the diode. The sensitivity change in V_T (V/mA) is such that appreciably lower power is required and the overall power gain is thereby increased. The unit may also be switched off by the application of a negative pulse to the p_+-junction.

(ii) *"Transistor" characteristic.* If the p_+-junction is operated as a transistor collector (negative bias) it will collect avalanche injected holes when the diode portion of the device is turned on. The avalanche dot then serves as an emitter with two stable states and the structure is similar to a transistor with a "built-in" memory. Experiments

with this type of structure have allowed the injection efficiency of an injecting avalanche to be estimated. The ratio of hole current to total current is found to have a maximum value of 20–30 per cent at low "on" currents and to decrease slowly with increasing current.

(iii) *High current-gain transistor.* If the avalanche dot is biased negatively to collect holes injected by the p_+-junction, the structure is essentially the same as that described by GRANVILLE[6], namely a transistor with high current gain. As shown by GRANVILLE, however, the minority carrier accumulation at the n_+-collector, which is the origin of the current gain, also severely limits the frequency response of this arrangement.

7. SUMMARY AND CONCLUSIONS

We have demonstrated the feasibility of making a two-terminal negative-resistance device employing the avalanche injection effect, and indicated how two-state circuits may be constructed using this device. The merits of avalanche injection diodes are as follows:

(i) *Simplicity of fabrication.* The properties of only one junction, made by conventional alloying techniques, need be controlled, and a small spacing between two junctions is not required for high-speed operation. This may be compared with a *p–n–p–n* diode where three junctions and two spacings between junctions are important.

Simplicity of fabrication implies high reproducibility and it should be noted that the histogram given in Fig. 5 does not indicate the reproducibility but merely the degree of agreement between a simplified theory and experiment.

(ii) *High speed.* Though the lowest turn-on times measured by us are about 3 mμsec, this figure is believed to be determined by the test equipment. Estimates by GUNN[1] indicate turn-on times of about 3 mμsec. Theoretically a turn-on time of the order of y_0/v_s (Section 4) would be expected, which is appreciably less than 1 mμsec.

(iii) *Independence of temperature.* This feature, discussed in Section 4, arises essentially from the absence of *p–n* junctions in the device. The high limiting temperature for germanium ($\sim$140°C) and even higher value for silicon ($\sim$350°C) allows very high power dissipation if an efficient heat sink is used.

The chief disadvantage of an avalanche injection diode, when compared with a *p–n–p–n*, is the high sustaining voltage. Fortunately the insensitivity of the device to temperature allows a high power dissipation in the on state and the power drain from the supplies is independent of V_s, so this feature is not unduly serious. Furthermore, V_s is so reproducible from unit to unit that it can, in some circuits, be "backed off" by a biased *p–n* junction diode.

Acknowledgements—For the supply of germanium and silicon crystals we are indebted to Mr. P. J. HOYLAND and Miss J. HUTCHEON respectively. Mr. J. L. CLARKE made a large number of the diodes and triodes, particularly silicon units, and Mr. A. C. BAYNHAM did much of the circuit work.

The paper is published by permission of the Controller, H.M. Stationery Office.

REFERENCES

1. J. B. GUNN, *Proc. Phys. Soc. Lond.* B**66,** 781 (1956).
2. E. J. RYDER, *Phys. Rev.* **90,** 766 (1953).
3. J. B. GUNN, *J. Electronics* **2,** 88 (1956).
4. A. F. GIBSON and J. R. MORGAN, *Brit. Pat. Appl.* 29800/57 (1957).
5. D. E. MASON and D. F. TAYLOR, *Progr. Semiconductors* **3,** 85 (1958).
6. A. F. GIBSON and J. W. GRANVILLE, *J. Electronics* **2,** 259 (1956).
7. M. A. TOWNSEND, *Bell Syst. Tech. J.* **36,** 755 (1957).
8. J. B. GUNN, *Progr. Semiconductors* **2,** 213 (1957).
9. K. G. MCKAY, *Phys. Rev.* **94,** 877 (1954).
10. S. L. MILLER, *Phys. Rev.* **99,** 1234 (1955).
11. W. SHOCKLEY, *Bell Syst. Tech. J.* **30,** 990 (1951).
12. C. HERRING, *Bell Syst. Tech. J.* **34,** 237 (1955).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 70–74. Printed in Great Britain

A METHOD OF MAKING NEGATIVE-RESISTANCE TWO-STATE SWITCHING DIODES

C. A. HOGARTH

Royal Radar Establishment, Great Malvern, Worcs.*

(*Received* 20 *June* 1959)

Abstract—It is shown that an assembly consisting of a point contact bearing on a surface made heavily *n*-type by diffusion of phosphorus into a 20 Ω-cm *n*-type silicon crystal, possesses a voltage–current characteristic typical of the avalanche injection diodes described by GUNN and by GIBSON and MORGAN.

Résumé—On démontre qu'un élément constitué d'un contact à pointe sur une surface dopée très fortement *n* par la diffusion de phosphore dans du silicium de type *n* à 20 Ω-cm possède une caractéristique courant–tension typique des diodes à avalanche d'injections décrites par GUNN et par GIBSON et MORGAN.

Zusammenfassung—Es wird gezeigt, dass eine Anordnung, welche aus einem Punktkontakt auf einer Oberfläche besteht, deren *n*-Leitung durch Eindiffusion von Phosphor in ein *n*-Typ-Siliziumkristall verstärkt wurde, eine Stromspannungskennlinie besitzt, die für Dioden mit Lawinen-Injektion typisch ist, wie sie sowohl von GUNN als auch von GIBSON und MORGAN beschrieben wurden.

1. INTRODUCTION

THERE is at present considerable interest in the development of fast negative-resistance diodes for use in many electronic circuits for switching purposes. Two main types have been developed, the *p–n–p–n* diode switch described by MOLL *et al.*[1] and the *L–H* junction operated in the avalanche injecting mode as described by GUNN[2] and by GIBSON and MORGAN[3]. In each of these devices when operated in the "forward" direction, a very small current is passed until the applied voltage reaches some critical value or "trigger" voltage. A sharp negative-resistance region is then reached, the voltage across the junction drops to a lower "sustaining" voltage and, on increasing the applied field, the current increases rapidly at an approximately constant value of sustaining voltage. GIBSON and MORGAN[3] have given a design theory of the *L–H* diode and have shown that the characteristics of devices, made by alloying small n_+-regions to *n*-type crystals followed by a controlled etching treatment to reduce the junction area, agree very well with the conclusions of this theory. It was considered that a different approach to the production of the *L–H* diode switch would be of interest and the investigation described here was carried out in order to demonstrate the feasibility of a quite different method of construction.

2. EXPERIMENTAL

(i) *Method*

The method adopted was to make use of two techniques, a modern one, impurity diffusion into silicon, and an old one, point-contact rectification. The principle was simply to use the diffusion technique to produce an n_+–n junction near the surface of a high quality silicon crystal and to make a point-contact connection to this surface layer as indicated in Fig. 1. Provided that the layer thickness t is small enough so that very little current flows along the surface layer, most of the current will flow across the n_+–n junction immediately below the point contact, and so long as the contact

* Now at the Brunel College of Technology, Acton, London W.3.

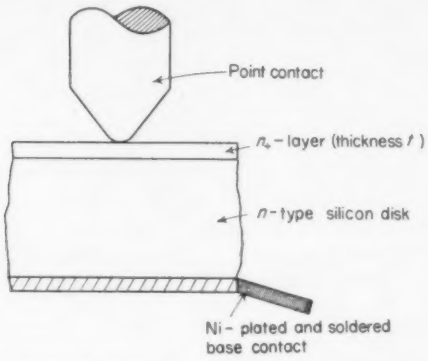


FIG. 1. Contact assembly for avalanche injection.

area is very small, a high electric field will be developed at the n_+–n interface and the conditions for avalanche injection will be satisfied.

ing with a 4 : 1 HNO_3/HF mixture for a few seconds, or

(b) by optically polishing with diamond pastes down to one containing $\frac{1}{2}$ μ particles.

The diffusion was carried out in a flowing-nitrogen ambient using analytical-reagent quality P_2O_5 as the source. A glassy deposit was laid down on the silicon at 900°C and the temperature was then raised to 1300°C for approximately 1 hr when appreciable diffusion of phosphorus took place.

After the diffusion process, experiments were carried out to determine the thickness of the n_+ layer produced. Because an n_+–n junction is very different in its electrical behaviour from the more usual p–n junction, the staining and etching techniques customarily employed for such junctions cannot be used. The method finally adopted was to make use of the very high surface concentration of

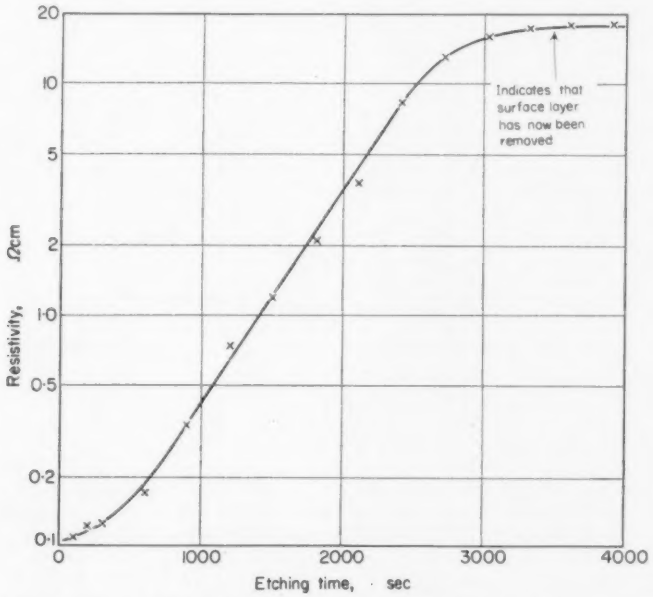


FIG. 2. Apparent resistivity of a silicon disk as a function of etching time.

(ii) *Material preparation*

The experiments were carried out using disks of 20 Ω-cm n-type silicon of thickness 0·5–1·0 mm. The surfaces were prepared in either of two ways prior to the diffusion,

(a) by lapping with 600 carborundum and etch-

impurities and thus the resultant high surface conductivity. The initial weight and thickness of the slice were measured, and then the apparent resistivity, using the four-probe machine. One side of the slice was painted with a polystyrene cement. The layer was etched slowly in steps using typically

a 50:1HNO_3/HF mixture and after each timed stage the resistivity ρ was again measured. It was found that on plotting ρ against time of etching the values initially remained low, increased fairly rapidly, and then levelled off at a value characteristic of the bulk material. A typical result is shown in Fig. 2. When the higher values of ρ are reached the polystyrene is dissolved off and the thickness and/or weight are also monitored. From the value at which the value of ρ levels off, a fairly precise value of layer thickness can be estimated.

voltage–current characteristics were observed on a cathode ray oscilloscope using a conventional test circuit. For the best results wire of diameter 0·002 in. was employed for the point contacts.

3. EXPERIMENTAL RESULTS

Voltage–current characteristics of some diode assemblies, made as described above, are shown in Figs. 3 and 4. Results for DS8 (etched before the diffusion process) in Fig. 3 show some of the characteristics of a point contact and how they change

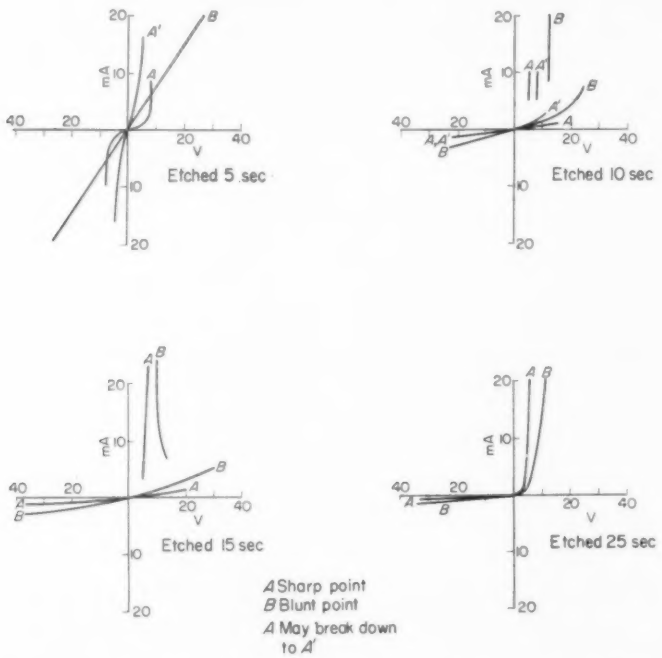


FIG. 3. Characteristics of avalanche diodes made on specimen DS8. (N.B. Etched before diffusion process.) (a), (b) and (c) show the effect of thinning the surface n_+-layer; (d) indicates normal rectification and the complete removal of the n_+-layer.

(iii) *Diode investigation*

Specimens about 4 mm square were used for the investigation. The side opposite the n_+-layer to be used was lapped with carborundum, nickel plated, and soldered to fairly massive brass bases to act as heat sinks. Contacts of electrolytically-pointed phosphor–bronze or tungsten were made to the surfaces by means of a micro-manipulator and the as the n_+-surface layer thickness is reduced and finally removed by etching with fast CP-4. To obtain the required characteristics, the optimum time of etching in fast CP-4, for surfaces which have had 1 hr of phosphorus diffusion at 1300°C, is between 15 and 20 sec. From resistivity measurements, made as described above, the resultant layer thickness was estimated as between

0·0001 and 0·0002 in. typically. Results for DS9 (optically polished before the diffusion process) in Fig. 4 again show that 15–20 sec is a suitable etching period for obtaining a suitable current–voltage characteristic. A considerably longer time is necessary before the layer is completely removed. In general, and particularly for contacts on surfaces which have been etched for the optimum period, the V–I characteristic is the same all over the slice. The results described here suggest that etching before the diffusion is to be preferred.

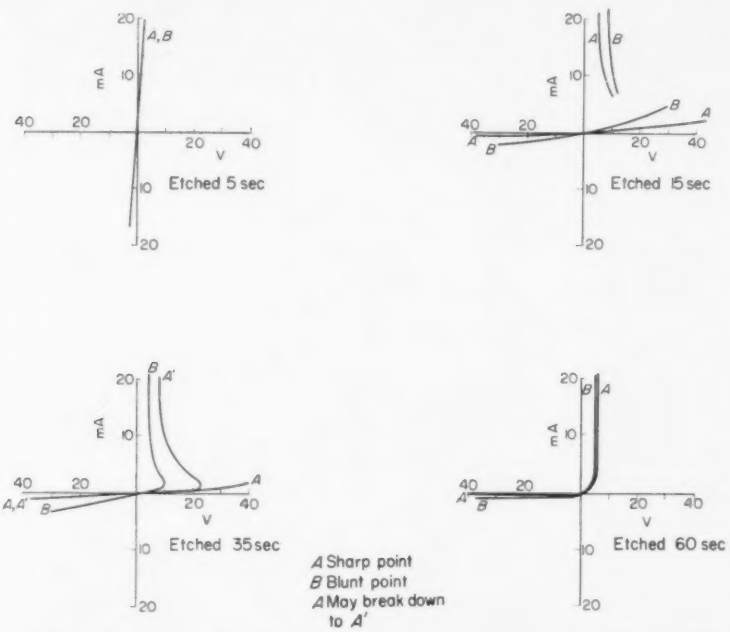


FIG. 4. Characteristics of avalanche diodes made on specimen DS9. (N.B. Polished before diffusion process.) (d) indicates that a longer period is required to etch off the n_+-layer, which forms in a "patchier" way, than for specimen DS8.

It may be noted that in each case the trigger voltage of the device is approximately 30 V and the current at this voltage is about 1 mA when a sharp point is used and 5 mA for a blunt point on thicker contact wire. The sustaining voltage is 5–7 V when a sharp point is used but 10–12 V for a blunt point. Currents drawn by the device are shown in the diagrams up to 20–30 mA but in practice currents of order 100 mA or more can be drawn without deterioration, provided that the heat-dissipation arrangements are adequate.

4. DISCUSSION

The feasibility of making negative-resistance diodes by means of a point contact to an etched high-conductivity diffused surface layer on silicon has been demonstrated. The general properties of the device are as predicted by the design theory of GIBSON and MORGAN[3] and it is interesting to note the similarity in behaviour between this device and one made by alloying. In particular it would appear that a lower sustaining voltage can be achieved by the present method provided sharp point contacts are used. The theory given by GIBSON and MORGAN[3] suggests that low values of the sustaining voltage can be obtained if the contact radius is less than a critical value of about 10^{-4} cm and it is believed that this condition is achieved with sharp points.

Devices made by the present method are expected to be especially useful if multicontact assemblies are used. The heat treatment of silicon during the diffusion process reduces the minority carrier lifetime to 1 μsec or less, so that the diffusion

length is $\sim 3 \times 10^{-3}$ cm and points could be placed on the surface at distances of 10^{-2} cm without fear of injection at one contact triggering a neighbouring one. In principle, the operation of the device depends only on the area of the point contact and the thickness of the surface layer. The main body of the silicon may be quite thick so that the silicon could, to some extent, act as its own heat sink.

Acknowledgements—It is a pleasure to acknowledge the help I have received in this work, notably from Dr. A. F. GIBSON, Mr. G. W. GREEN, who grew the silicon crystals used, and Miss JEAN HUTCHEON who assisted in the experimental work.

The paper is published by permission of the Controller, H.M. Stationery Office.

REFERENCES

1. J. L. MOLL, M. TANENBAUM, J. M. GOLDEY and N. HOLONYAK, *Proc. Instn. Radio Engrs. Aust.* **44,** 1174 (1956).
2. J. B. GUNN, *Proc. Phys. Soc. Lond.* **69**B, 781 (1956).
3. A. F. GIBSON and J. R. MORGAN, *Solid-State Electron.* **1,** 54 (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 75–84. Printed in Great Britain

HALLEFFEKT-VIERPOLE MIT HOHEM WIRKUNGSGRAD

G. ARLT

Allgemeine Deutsche Philips Industrie Gmb H., Zentrallaboratorium,
Laboratorium Aachen, Aachen, Germany

(*Received* 10 *August* 1959)

Zusammenfassung—Der Hallgyrator mit 2 Steuerelektroden und 2 Hallelektroden ist ein Vierpol mit einem Wirkungsgrad $<0,17$. Durch Verwendung mehrerer Steuer- und Hallkreise an einem Hallplättchen lässt sich der Wirkungsgrad wesentlich verbessern. Es werden Methoden zur Berechnung derartiger Hallvierpole angegeben und Vorschläge für Ausführungsformen im Nieder-, Hoch- und Ultrahochfrequenzgebiet gegeben. Die Erhöhung des Wirkungsgrades gilt in gleicher Weise für alle Halleffektanwendungen: Generator, Gyrator, Isolator, Circulator, etc.

Durch Reihenparallelschaltung eines Hallgyrators mit einem Transformator lässt sich Einwegleitung mit geringer Vorwärtsdämpfung erreichen.

Abstract—The upper efficiency limit of the Hall-Gyrator with four electrodes is 0·17. To enlarge the efficiency one has to solder more than one polarizing-current circuit and Hall circuit to the same Hall specimen. This holds for all types of Hall devices such as generator, gyrator, isolator, circulator, etc. Methods are given for the calculation of the highest possible efficiency and other interesting properties. Gyrators and isolators are proposed for audio, high and ultra-high frequency applications.

Series–parallel connexion of gyrator and transformator yields a low-loss uniline.

Résumé—Le gyrateur à effet Hall avec deux électrodes de polarisation et deux électrodes Hall est un quadripole avec un rendement inférieur à 0,17. On peut améliorer considérablement ce rendement en augmentant dans le même échantillon le nombre des circuits de polarisation et des circuits de Hall. On donne des méthodes pour calculer de tels quadripoles à effet Hall et pour déterminer leur géométrie pour des basses, hautes et très hautes fréquences. L'augmentation du rendement est également valable pour tous les systèmes à effet Hall: générateur, gyrateur, isolateur, circulateur, etc.

Par un branchement série–parallèle d'un gyrateur à effet Hall avec un transformateur on obtient une ligne unidirectionnelle à faible perte.

EINLEITUNG

Die üblichen Hallgeneratoren[1–3] nach Abb. 1(a, b) benutzen Elektroden aus einem Material hoher Leitfähigkeit, welches kurzschliessend auf die Hallspannungen oder sonstige Potentialunterschiede in der Nähe der Elektroden wirkt. Die dabei entstehenden Kreisströme führen zu Leistungsverlusten in der Hallprobe, durch die sowohl das Verhältnis von Ausgangs- zu Eingangsspannung als auch der Energieübertragungswirkungsgrad nach oben begrenzt werden. Zur Erhöhung des Spannungsverhältnisses hat Perrier[4] eine Schaltung vorgeschlagen, bei der der Eingangsstromkreis in eine Vielzahl von parallel liegenden Eingangsstromkreisen aufgespalten wird. Diese Schaltung gleicht einer Serienschaltung verschiedener Hallplättchen auf der Ausgangsseite und einer eingangsseitigen Parallelschaltung derselben. Dieses Prinzip lässt sich in allgemeiner Form erfassen und erweitern.

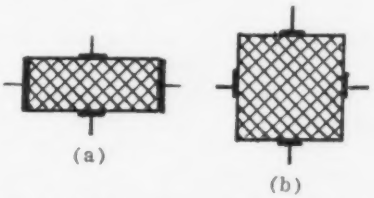


Abb. 1. Hallgyratoren mit 4 Elektroden.

1. DER HALLVIERPOL MIT ANISOTROPEN ZULEITUNGEN

Das rechteckige dünne Hallplättchen der Abb. 2 werde von einem Magnetfeld senkrecht durchdrungen. An den Kanten, an denen bisher wie in Abb. 1(a, b) nur jeweils eine Elektrode befestigt

war, sind nun mehrere Elektroden und Steuerkreise 1–1′ bis 7–7′ und die Elektroden mit den Verbraucherkreisen *a–a′* bis *f–f′* angebracht. Entsteht z. B. zwischen den Elektroden 2′ und 5′ eine Hallspannung, so ist diese nun nicht mehr wie bei

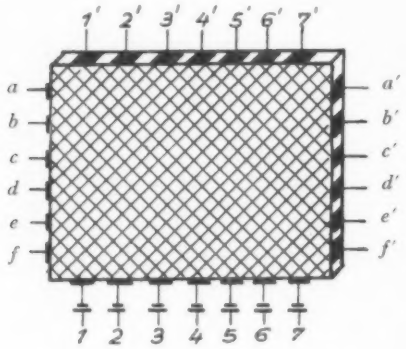


Abb. 2. Vielelektroden-Hallgyrator.

den kompakten Elektroden der Abb. 1 kurzgeschlossen, denn die Steuerstromkreise 2–2′ und 5–5′ stehen ausser über das Hallplättchen in keiner elektrisch leitenden Verbindung. Sämtliche Elektroden können auf verschiedenem Potential liegen. Durch eine derartige Schaltungsanordnung wird man weitgehend von den üblichen Geometriebedingungen für Hallplättchen frei. Zu einer theoretischen Untersuchung wählt man am besten den Sonderfall, bei dem Steuer- u. Verbraucherkreis aus anisotropen Elektroden, Zuleitungen, Spannungsquellen und Widerständen bestehen, in dem Sinne, als ob die Zahl der getrennten, voneinander isolierten Kreise gegen unendlich gehe. Abb. 3 zeigt eine derartige Schaltung aus einem parallelogrammförmigen Hallplättchen A von geringer Dicke d, dessen Winkel um die Grösse α von einem rechten Winkel abweichen. An dieses Plättchen sind Zuleitungsbänder B und C angeschlossen, die anisotrope Leitfähigkeitseigenschaften mit den Leitfähigkeitstensoren

$$(\sigma)_B = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_B \end{pmatrix}; \qquad (\sigma)_C = \begin{pmatrix} \sigma_C & 0 \\ 0 & 0 \end{pmatrix} \tag{1a, b}$$

haben. Sieht man die Leitung C als Verbraucherstromkreis an, so kann man z.B. durch deren Länge l_C und die Grösse σ_C den Aussenwiderstand festlegen. In den Steuerstromkreis B sei eine anisotrope Spannungsquelle mit der Leerlaufspannung U und dem Innenwiderstand Null eingeschlossen, die Tensorkomponente σ_B gehe gegen ∞. Für das Hallplättchen A lautet der Leitfähigkeitstensor

$$(\sigma)_A = \begin{pmatrix} +\sigma_{11} & +\sigma_{12} \\ -\sigma_{12} & +\sigma_{11} \end{pmatrix} \tag{1c}$$

mit $$\sigma_{11} = \frac{\sigma_0}{1+\beta^2}, \quad \sigma_{12} = \frac{\sigma_0 \beta}{1+\beta^2}$$

und $$\beta = \sigma_0 R B,$$

wobei σ_0 die Leitfähigkeit *, R die Hallkonstante des Halbleitermaterials und B die magnetische Induktion senkrecht zum Hallplättchen darstellen. Eine derartige Anordnung nach Abb. 3, die sich praktisch nur näherungsweise verwirklichen lässt, entspricht der Abb. 2 mit unendlich dicht nebeneinander liegenden Elektroden und Leitungen.

Für das Modell der Abb. 3 lassen sich die Feldgleichungen unter Berücksichtigung der Randbedingungen elementar lösen. Es zeigt sich, dass

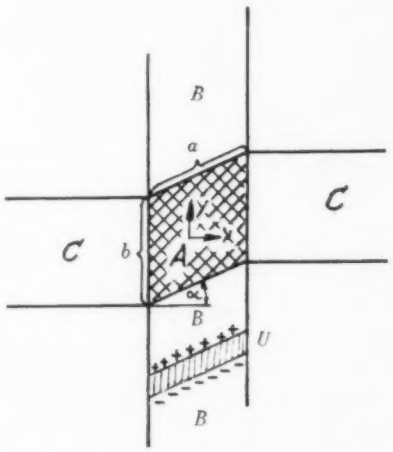


Abb. 3. Schiefer Hallvierpol mit anisotropen Elektroden, Zuleitungen und anisotroper Spannungsquelle.

die Äquipotentiallinien im Hallplättchen Geraden sind, so dass die elektrische Feldstärke ortsunabhängig ist. Für die Feldstärkekomponenten E_x^A

* Genau stellt σ_0 die Leitfähigkeit eines sehr langen dünnen Stabes aus dem Halbleitermaterial im Felde B dar.

und E_y^A im Plättchen A ergibt sich

$$E_x^A = -\frac{\sigma_{12}+\dfrac{\sigma_C a}{l_C}\sin\alpha}{\sigma_{11}+\dfrac{\sigma_C a}{l_C}\cos\alpha}\cdot\frac{U}{b}, \qquad E_y^A = \frac{U}{b} \quad (2a,b)$$

und für die Feldstärken in den Zuleitungen B und C

$$E_x^C = -\frac{a}{l_C}(E_x^A\cos\alpha + E_y^A\sin\alpha), \quad E_y^C = \frac{U}{b}, \quad (3a,b)$$

$$E_x^B = E_x^A + E_y^A \operatorname{tg}\alpha, \qquad E_y^B = 0. \quad (4a,b)$$

Die entsprechenden Stromdichten folgen aus der Tensorgleichung $\mathbf{i} = (\sigma)\,\mathbf{E}$ mit (σ) = Leitfähigkeitstensor. Man kann einfach nachweisen, dass diese Lösung sowohl die Potentialgleichung[2] $\Delta\varphi = 0$ als auch die Randbedingungen, nämlich die Stetigkeitsforderung für die Tangentialkomponente der elektrischen Feldstärke und für die Normalkomponente der Stromdichte erfüllt.

Aus der Feldlösung lässt sich ein formal der Vierpolgleichung entsprechender Ausdruck ableiten. Integriert man die Stromdichten i_x^C bzw. i_y^B über den Querschnitt der anisotropen Zuleitungen, so erhält man die Gesamtströme für Eingang und Ausgang:

$$J_1 = -i_y^B da\cos\alpha, \qquad J_2 = +i_x^C db, \quad (5a,b)$$

mit Vorzeichendefinitionen entsprechend Abb.4. Da die Potentialdifferenzen zwischen äquivalenten Punkten gegenüberliegender Kanten überall gleich sind, lässt sich auch eine Eingangs- und Ausgangsspannung definieren:

$$U_1 = -bE_y^A, \qquad U_2 = +l_C E_x^C. \quad (6a,b)$$

Man erhält auf diese Weise aus den Feldgleichungen (2–4) und den entsprechenden Stromdichtegleichungen sowie (1c) die formalen Vierpolgleichungen für den schiefen Hallgenerator mit anisotropen Elektroden

$$\begin{aligned} U_1 &= \frac{b}{ad\sigma_0\cos\alpha}J_1 - \frac{(\beta+\operatorname{tg}\alpha)}{\sigma_0 d}J_2, \\ U_2 &= -\frac{(\beta-\operatorname{tg}\alpha)}{\sigma_0 d}J_1 - \frac{a}{bd\sigma_0\cos\alpha}J_2. \end{aligned} \quad (7)$$

Die Leerlaufübertragungswiderstände in (7) für

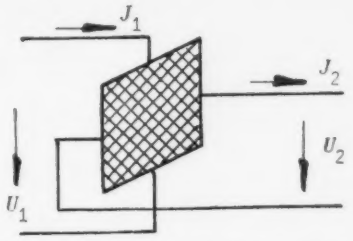


Abb. 4. Definition der Vorzeichen von Strömen und Spannungen.

Vorwärtsbetrieb

$$R_{21} = -\frac{(\beta-\operatorname{tg}\alpha)}{\sigma_0 d} = R_a + R_r$$

und für Rückwärtsbetrieb

$$R_{12} = -\frac{(\beta+\operatorname{tg}\alpha)}{\sigma_0 d} = R_a - R_r$$

bestehen aus einem reziproken Anteil

$$R_r = tg\alpha/\sigma_0 d,$$

der aus der Unsymmetrie des Plättchens herzuleiten ist und in erster Näherung* magnetfeldunabhängig ist, und einem antireziproken Anteil $R_a = -\beta/\sigma_0 d$, welcher allein auf Grund des Halleffektes zustande kommt. Bei einem rechteckigen Hallplättchen, $\alpha = 0$, entfällt der reziproke Anteil. Der Vierpol stellt dann einen Hallgyrator dar. Wenn die Absolutwerte beider Anteile gleich sind $R_r = \pm R_a$, so verschwindet einer von den beiden Leerlaufübertragungswiderständen R_{21} bzw. R_{12}. Der Vierpol wird damit zum Einwegleiter oder Isolator.

Mittels der Vierpolgleichungen (7) lassen sich in bekannter Weise die Spannungsübersetzung U_2/U_1, die Stromübersetzung J_2/J_1, die Übertragungswiderstände U_1/J_2 und U_2/J_1 sowie Eingangswiderstand U_1/J_1 und Wirkungsgrad U_2J_2/U_1J_1 für beliebige Aussenwiderstände R_a berechnen.

Linear in der magnetischen Induktion B bzw. in β sind nur J_2/J_1 sowie U_2/J_1. Schreibt man den Abschlusswiderstand in Einheiten von $a/bd\sigma_0$ also $R_a = a\lambda/bd\sigma_0$, so erhält man für den Eingangs-

* Vgl. Anmerkung auf S. 76.

widerstand $R_e = U_1/J_1$:

$$R_e = \frac{b}{ad\sigma_0} \cdot \left(\frac{\frac{\lambda}{\cos\alpha} + 1 + \beta^2}{\lambda + \frac{1}{\cos\alpha}} \right) \tag{8}$$

und für den Wirkungsgrad $\eta = U_2 J_2 / U_1 J_1$:

$$\eta = \frac{\lambda(\mathrm{tg}\alpha - \beta)^2}{\left(\frac{\lambda}{\cos\alpha} + 1 + \beta^2\right)\left(\lambda + \frac{1}{\cos\alpha}\right)}. \tag{9}$$

Maximale Leistung wird übertragen, wenn

$$\lambda = \sqrt{(1+\beta^2)},$$

also

$$R_{a\,\mathrm{opt}} = \frac{a}{bd\sigma_0}\sqrt{(1+\beta^2)}. \tag{10}$$

Damit ist der maximale Wirkungsgrad

$$\eta_{\max} = \frac{\sqrt{(1+\beta^2)} - \frac{1}{\cos\alpha}}{\sqrt{(1+\beta^2)} + \frac{1}{\cos\alpha}} \cdot \frac{\beta - \mathrm{tg}\alpha}{\beta + \mathrm{tg}\alpha}. \tag{11}$$

Ähnliche Ausdrücke erhält man für den umgekehrten Vierpol, bei dem also Eingang und Ausgang vertauscht sind. So lauten der optimale Anpassungswiderstand $R'_{a\mathrm{opt}}$ und der maximale Wirkungsgrad $\eta'_{\max}$ in diesem Falle

$$R'_{a\,\mathrm{opt}} = \frac{b}{ad\sigma_0}\sqrt{(1+\beta^2)}, \tag{12}$$

$$\eta'_{\max} = \frac{\sqrt{(1+\beta^2)} - \frac{1}{\cos\alpha}}{\sqrt{(1+\beta^2)} + \frac{1}{\cos\alpha}} \cdot \frac{\beta + \mathrm{tg}\alpha}{\beta - \mathrm{tg}\alpha}. \tag{13}$$

Hallgyrator

Für den rechteckigen Hallgenerator mit anisotropen Zuleitungen ist $\alpha = 0$ und daher der maximale Wirkungsgrad nach (11)

$$\eta_{\max} = \frac{\sqrt{(1+\beta^2)} - 1}{\sqrt{(1+\beta^2)} + 1}. \tag{14}$$

β ist für Störstellenhalbleiter das Produkt aus Hallbeweglichkeit und magnetischer Induktion. Für Halbleiter hoher Trägerbeweglichkeit, z.B. InSb kann β erheblich grösser als 1 werden und somit lassen sich nach (14) Wirkungsgrade sehr nahe an 1 erreichen. Im Gegensatz zu dem Hallgenerator mit 4 Elektroden[2] (Abb. 1) besteht also keine prinzipielle obere Grenze für den Wirkungsgrad. Ebenso sind an die geometrischen Formen des Hallplättchens nicht mehr die Bedingungen geknüpft, die man für gute Hallgeneratoren bzw. Gyratoren mit 4 Elektroden fordern muss. Bei den üblichen 4-Elektroden Hallgeneratoren bewirken Grösse und Form der Hallsonden und der Halbleiterplatte eine Abweichung von dem linearen Zusammenhang zwischen Hallspannung und magnetischer Induktion.[3] Auch dieser, im allgemeinen unerwünschte Effekt, fehlt bei dem Hallgenerator mit anisotropen Elektroden. Eine geringe Störung der Linearität zwischen Hallspannung bzw.-strom und magnetischer Induktion kann allerdings noch über die Magnetfeldabhängigkeit der Ladungsträgerbeweglichkeit erfolgen, die in jedem Koeffizienten der Vierpolgleichung (7) in den Grössen σ_0 und β indirekt enthalten ist (Vgl. Anmerkung auf S.76).

Hallisolator

Ist das Plättchen parallelogrammförmig und stellt man das Magnetfeld so ein, dass $\beta = -\mathrm{tg}\alpha$, so findet in der Vorwärtsrichtung nach (9) eine Energieübertragung mit dem Wirkungsgrad

$$\eta = \frac{4\lambda \cos\alpha \,\mathrm{tg}^2\alpha}{\left(\lambda + \frac{1}{\cos\alpha}\right)^2}, \text{ bzw. } \eta_{\max} = \mathrm{sin}.^2\alpha \tag{15, 16}$$

statt, während in der umgekehrten Richtung nach (11) kein Energiefluss durch das Hallplättchen erfolgt: $\eta' = \eta'_{\max} = 0$. Auch in diesem Falle des Isolators gibt es keine prinzipielle Grenze für die Dämpfung in Durchlassrichtung.

2. HALLVIERPOLE MIT ENDLICHER ELEKTRODENZAHL

(a) *Eigenschaften*

Ideale anisotrope Leiter und Widerstände, wie sie für das Modell des vorigen Abschnitts vorausgesetzt wurden, gibt es nicht. Man muss sich mit Näherungslösungen z.B. in der Art nach Abb.2

begnügen. Je grösser die Zahl der voneinander getrennten Kreise ist, desto näher wird man dem Ergebnis des vorigen Abschnitts kommen.

In diesem Abschnitt sollen die theoretischen Grenzwerte von optimalem Anpassungswiderstand und maximalem Wirkungsrad für Gyratoren und Isolatoren mit endlicher Elektrodenzahl bestimmt werden, für den Fall, dass das Produkt aus Trägerbeweglichkeit und magnetischer Induktion μB gegen unendlich bzw. der Hallwinkel gegen 90° geht.

WICK[2] hat an dem Vier-Elektroden-Gyrator nach Abb. 1(a,b) mit Hilfe eines Satzes der Potentialtheorie gezeigt, dass die Potentiale an den Hallkontakten auch bei beliebig hohen endlichen magnetischen Feldstärken keine Extremwerte sein können. Die Extrema sind immer an den stromführenden Steuerelektroden zu finden, die mit einer Spannungsquelle verbunden sind. Lässt man den Hallwinkel gegen 90° streben,* so werden Stromlinien und Äquipotentiallinien immer stärker deckungsgleich. Das bedeutet, dass eine Randlinie, die immer auch Stromlinie ist, im Grenzübergang auch zur Äquipotentiallinie wird. Die Potentiale beliebiger Randkontakte (bei einfach zusammenhängenden Flächen) streben also bei wachsendem Produkt μB gegen die an den Steuerelektroden vorhandenen Extremwerte.

Die Erweiterung dieses Gedankens auf Hallplättchen mit mehreren Steuerkreisen, soll an dem dünnen, beliebig geformten Hallplättchen der Abb. 5 gezeigt werden.

Liegt nur im Kreise 1, also zwischen den Kontakten 1 und 1′, eine Spannungsquelle U_1, so können die Potentiale φ_2, φ'_2 bis φ_m, φ'_m an den Kontakten 2–2′ bis m–m′ nur im Bereich zwischen 0 und U_1 liegen. Im Grenzübergang nach dem Hallwinkel 90° streben die Potentiale aller Randkontakte nach den Potentialen 0 oder $U_1 = \varphi_1$ je nach der geometrischen Lage. Wird die Spannungsquelle in Kreis 1 entfernt und eine Spannungsquelle U_2 in Kreis 2 eingefügt, so gilt entsprechend, dass die Potentiale aller Kontakte ausser denen des Kreises

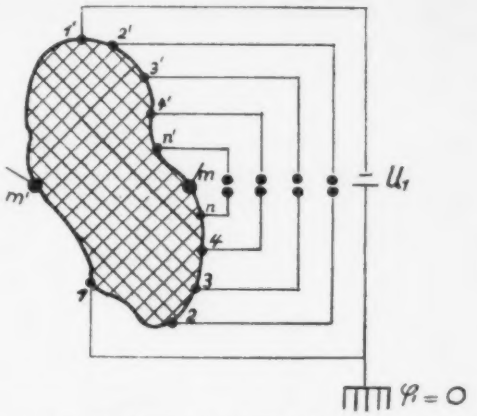


ABB. 5. Zur Berechnung der maximalen Hallspannung (siehe Text).

2 im Bereich zwischen φ_2 und φ'_2 liegen müssen bzw. im Grenzübergang diese Werte gerade erreichen. Dieses Verfahren lässt sich für die Kreise 3,4 bis n fortsetzen. Superponiert man schliesslich die verschiedenen Lösungen, d.h. fügt man in jeden der Kreise 1 bis n eine Spannungsquelle U_l ein, so kann durch Addition der Einzelpotentiale zwischen zwei stromlosen Randkontakten m–m' nur eine Potentialdifferenz $\varphi_m - \varphi'_m = U_m$ entstehen, deren Absolutbetrag kleiner ist als die Summe aller Einzelspannungen U_l bzw. im Grenzübergang dieser Summe gleich wird:

$$|U_m| < |\Sigma U_l|, \qquad \lim_{\mu B \to \infty} |U_m| = |\Sigma U_l|. \tag{17a, b}$$

Ein Netzwerk, wie in Abb. 2 dargestellt, lässt sich ähnlich wie im vorigen Abschnitt wieder als Vierpol betrachten, wenn man die Ströme in allen Verbraucherkreisen zu einem Gesamtstrom J_2 und die Einzelströme in allen Steuerkreisen zu einem Gesamtsteuerstrom J_1 addiert. Wir setzen weiter voraus, dass alle Steuerkreise unter sich und alle Verbraucherkreise unter sich gleich sind. Dann erhält man für den Gyrator die Vierpolgleichungen

$$\begin{aligned} U_1 &= R_{11}J_1 - R_{12}J_2, \\ U_2 &= -R_{12}J_1 - R_{22}J_2. \end{aligned} \tag{18}$$

Führt man $\epsilon^2 = R^2_{12}/R_{11}R_{22}$ ein[3], so muss man für maximale Leistungsübertragung einen Aussenwiderstand

$$R_{a\,\mathrm{opt}} = R_{22}\sqrt{(\epsilon^2+1)} \tag{19}$$

* Ob dieser Grenzwert überhaupt angenähert werden kann, ist eine andere Frage. Selbst für unendlich grosse Magnetfelder ist es möglich, dass das Produkt μB einen endlichen Wert nicht überschreitet, dann nämlich, wenn die Beweglichkeit, die ja selbst eine Fuktion von B ist, in starken Magnetfeldern proportional zu $1/B$ oder einer höheren Potenz davon ist.

fordern. Als maximaler Wirkungsgrad ergibt sich dann

$$\eta_{\max} = \frac{\sqrt{(\epsilon^2+1)}-1}{\sqrt{(\epsilon^2+1)}+1} \tag{20}$$

Besteht der Vierpol aus einem Hallplättchen mit m gleichartigen Verbraucherkreisen und n gleichartigen Steuerkreisen, so ist nach den obigen Forderungen (17) für die Hallspannung

$$|U_2| < n|U_1|, \text{ bzw. } \lim_{\mu B \to \infty} |U_2| = n|U_1|, \tag{21a, b}$$

und daraus

$$\epsilon^2 < nm, \text{ bzw. } \lim_{\mu B \to \infty} \epsilon^2 = nm. \tag{25a, b}$$

Die Leistungsübertragung ist dann nach (20) und (25) beschränkt auf

$$\eta_{\max} < \frac{\sqrt{(nm+1)}-1}{\sqrt{(nm+1)}+1}. \tag{26}$$

Für den quadratischen Gyrator mit gleicher Anzahl von Steuer- und Verbraucherkontakten $n = m$ ergeben sich daraus obere Grenzen (vgl. jedoch Anmerkung auf S.76) für den Wirkungsgrad nach Tabelle 1.

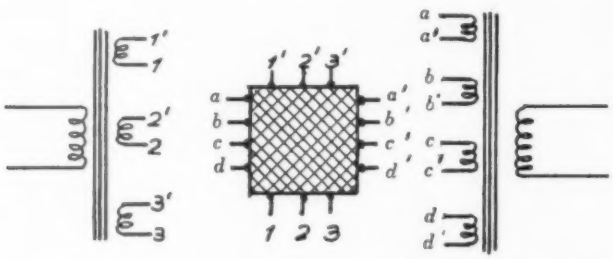


ABB. 6. Transformatorspeisung des Vielelektroden-Hallgyrators.

oder wenn Eingang und Ausgang vertauscht werden

$$|U'_1| < m|U'_2|, \text{ bzw. } \lim_{\mu B \to \infty} |U'_1| = m|U'_2|. \tag{22a, b}$$

Aus diesen Bedingungen ergibt sich mit Hilfe der Vierpolgleichungen (18)

$$\frac{|R_{12}|}{R_{11}} < n, \qquad \lim_{\mu B \to \infty} \frac{|R_{12}|}{R_{11}} = n, \tag{23a, b}$$

$$\frac{|R_{12}|}{R_{22}} < m, \qquad \lim_{\mu B \to \infty} \frac{|R_{12}|}{R_{22}} = m, \tag{24a, b}^*$$

* Aus (23b) und (24b) folgt

$$\lim \frac{R_{22}}{R_{11}} = \frac{n}{m}$$

ein Ergebnis, das von der Geometrie des Hallplättchens unabhängig ist! Das bei endlichen Feldstärken durch die Geometrie des Hallplättchens, durch die Form der Kontakte, die Anzahl der Kreise und durch die Magnetfeldstärke bestimmte Verhältnis der Leerlaufwiderstände R_{22}/R_{11}, wird im Grenzübergang nur noch eine Funktion der Anzahl der Kreise. Beide Leerlaufwiderstände R_{11} und R_{22} gehen im Grenzübergang nach ∞.

Tabelle 1.

$n =$	1	2	3	4	5	6	7
$\eta_{\max} <$	0,17	0,38	0,52	0,61	0,67	0,72	0,75

(b) *Speisung*

Zur homogenen Speisung der n Eingangskreise und zur gesammelten Energieentnahme an den m Ausgangskreisen gibt es verschiedene Methoden.

(i) Die n Eingangskreise werden über einen Transformator von einer Quelle gespeist, und die Energie wird an den Ausgangskreisen auf die gleiche Weise entnommen (Abb. 6). Eine sehr wirkungsvolle Abart dieser Schaltung, bei der Eingang oder Ausgang ideal anisotrop angeschlossen sind, besteht darin, das Hallplättchen als Hohlzylinder um den Kern des Transformators zu legen[5] (Abb. 7). Das den Halleffekt bewirkende magnetische Feld B muss dann den Zylinder radial durchstossen (in der Abb. 7 nicht dargestellt).

Die an den Hohlzylinderrändern angebrachten Kontaktgruppen können z.B. wieder über einen Transformator angeschlossen werden.

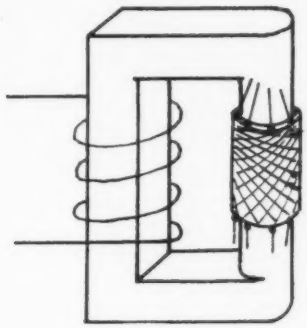


ABB. 7. Induktiv gespeistes, ringförmiges Hallplättchen.

(ii) Die einzelnen Kreise können auch galvanisch miteinander verbunden sein, aber induktiv getrennt werden wie in Abb. 8 dargestellt.

Eine für die Anwendung von Hallgyratoren in Hochfrequenzleitungen besonders günstige Art dieser Speisung besteht darin, die Parallelbandleitung oder das Koaxialkabel auf eine Länge $\lambda/4$ vom Plättchen aufzuschlitzen. λ sei die Wellenlänge der Hochfrequenz auf der Leitung. Abb. 9 zeigt ein rechteckiges Hallplättchen mit 4 Eingangskontaktpaaren und sechs Ausgangskontaktpaaren, an welche jeweils Bandleitungen angeschlossen sind, die auf eine Länge von $\lambda/4$ oder einem ungeradzahligen Vielfachen davon entsprechend der Kontaktzahl mit Schlitzen versehen sind. Die Zeichnung ist nicht masstabsgetreu. Die Plättchendimensionen müssen klein gegen die Wellenlänge sein. Das senkrecht zum Plättchen stehende Magnetfeld ist in Abb. 9 nicht eingezeichnet. Durch

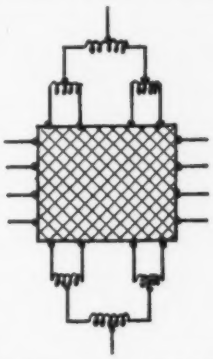


ABB. 8. Drosselspeisung des Vielelektroden-Hallgyrators.

die Schlitzung sind jeweils 2 Kontakte desselben Randes über $\lambda/4$ Stichleitungen, die einen für die betreffende Frequenz unendlich hohen Widerstand besitzen, verbunden.

Abb. 10 zeigt ein hohlzylindrisches Hallplättchen von einer Länge L, die klein gegen $\lambda/4$ ist, welches in den Aussenleiter eines Koaxialkabels eingebaut ist. Aussen- und Innenleiter des Kabels

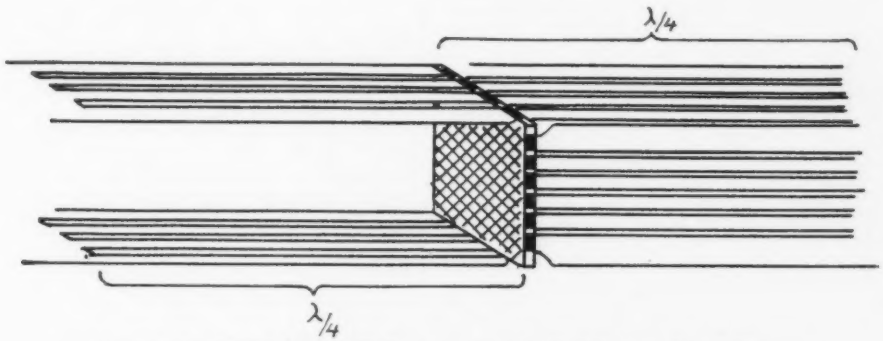


ABB. 9. Der Vielelektroden-Hallgyrator in einer Hochfrequenzbandleitung.

sind auf eine Länge $\lambda/4$—vom Plättchen aus gemessen—in Teilleitungen geschlitzt. Als Abschluss der Leitung dient z.B. eine geschlitzte speichenradförmige Kurzschlussscheibe, welche die Kontaktpunkte am endseitigen Rand des hohlzylindrischen Halbleiters mit den Teilleitungen des

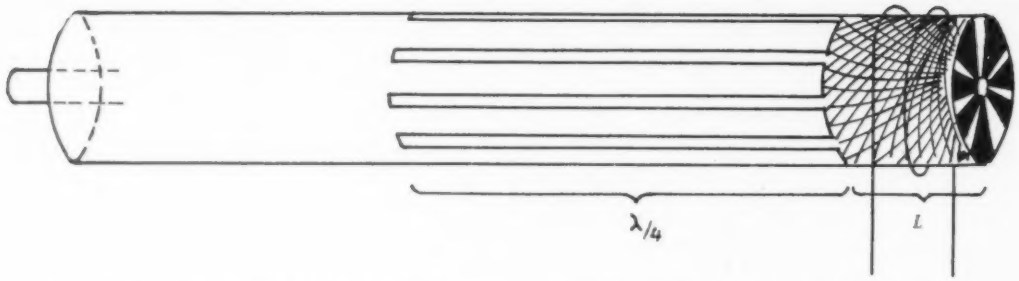


ABB. 10. Ringförmiger Vielelektroden-Hallgyrator in einem Koaxial-Kabel mit induktiver Kopplung.

geschlitzten Innenleiters verbindet. Durch das (nicht eingezeichnete) radiale Magnetfeld werden die das Plättchen durchfliessenden Ströme aufgeteilt in einen achsialen und einen zirkularen Anteil. Der zirkulare Anteil lässt sich induktiv ähnlich wie

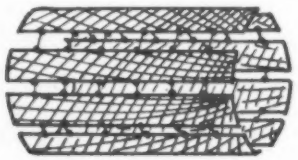


ABB. 11. Vereinfachter Aufbau des ringförmigen Vielelektroden-Hallgyrators.

in Abb. 7 also wieder ideal anisotrop aus dem Zylinder auf eine Paralleldrahtleitung auskoppeln. Die induktive Auskopplung kann durch eine galvanische ersetzt werden, indem der Hallzylinder einen Längsschlitz erhält, an welchem über mehrere Teilelektroden eine auf $\lambda/4$ geschlitzte Bandleitung angeschlossen wird.

Eingang und Ausgang sind vertauschbar. Der Vierpol stellt eine Symmetriereinrichtung über ein Hallplättchen dar und hat gyratorische Eigenschaften. Über die Wirkung des elektrisch steuerbaren radialen Magnetfeldes ist eine Modulation der Hochfrequenz möglich. Die Schwierigkeit, hohlzylinderförmige Hallplättchen herzustellen, lässt sich umgehen, indem man den Zylinder aus mehreren ebenen Hallplättchen zusammensetzt, die jeweils durch mehrere Kontakte, wie in Abb. 11 dargestellt, verbunden sind.

(c) *Isolator*

Entsprechend den Ergebnissen in Abschnitt 1 am Hallplättchen mit anisotropen Zuleitungen hat das parallelogrammförmige Plättchen bei Einhaltung der Beziehung $\operatorname{tg} \alpha = \mu B$ die Eigenschaften eines Einwegvierpols. In derselben Weise, wie im vorigen Absatz an das rechteckige Hallplättchen auf $\lambda/4$ geschlitzte Bandkabel ange-

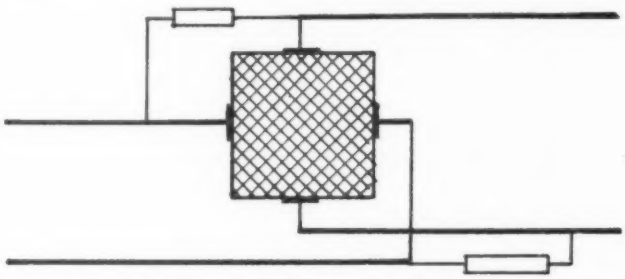


ABB. 12. Isolatorschaltung des Hallvierpols durch Parallelwiderstände.

schlossen wurden, lassen sich auch an das parallelogrammförmige Plättchen geschlitzte Bandkabel anschliessen, um einen für Meter und Dezimeterwellen geeigneten Einwegleiter zu erhalten.

In ähnlicher Weise lässt sich auch eine Einwegleitung mit Übergang vom Koaxialkabel auf

eine symmetrische Leitung (bzw. umgekehrt) herstellen, indem nämlich der nach Abb. 10 im geschlitzten Koaxialkabel befindliche hohlzylindrische Hallring mit einem spiralförmigen Schlitz versehen wird, so dass bei einem gedachten Abrollen des Zylinders wieder ein Parallelogramm mit dem charakteristischen Winkel α entsteht. Wird an diesem spiralförmigen Schlitz ein Bandkabel angeschlossen, welches wieder durch $\lambda/4$ lange Schlitze die einzelnen Kontaktstellen entkoppelt, so erhält man einen symmetrierenden Einwegvierpol.

3. EINWEGLEITER DURCH TRANSFORMATOR-SCHALTUNG

Der normale Hallgyrator nach Abb. 1 mit vier Elektroden wird durch Parallelschaltung von Widerständen geeigneter Grösse nach Abb. 12 zum Einwegleiter.[1] Der Grenzwert des Wirkungsgrades in Durchlassrichtung für unendlich hohes Produkt μB ist

$$\eta_1 = 0{,}25$$

Für den schiefen Gyrator nach WICK[2] mit einer derartigen Symmetrie, dass in den Vierpolgleichungen $R_{11} = R_{22}$ ist, erhält man ebenfalls Wirkungsgrade, die den Wert 0,25 nicht überschreiten können.

Durch Reihenparallelschaltung von Transformator und Gyrator entsteht ein Einwegleiter, dessen Wirkungsgrad auch bei Verwendung des Vierelektroden-Gyrators wesentlich höher liegen kann. Diese Isolatorschaltung ist jedoch nur dann möglich, wenn der Gyrator verlustbehaftet ist, wenn also die Glieder R_{11} und R_{22} in den Vierpolgleichungen des Gyrators nicht verschwinden.

Geht man von den Vierpolgleichungen des Gyrators (18) aus und setzt für den verlustlosen Transformator

$$\begin{aligned} U_1 &= j\omega L_1 J_1 - j\omega M J_2, \\ U_2 &= j\omega M J_1 - j\omega L_2 J_2, \end{aligned} \tag{27}$$

mit L_1, L_2 = Primär- und Sekundärinduktivität, M = Gegeninduktivität, so erhält man durch Addition der Reihen-Parallel-Matrixelemente beider Gleichungssysteme folgende Reihenparallelform der Vierpolgleichungen für die Schaltung nach Abb. 13

$$\begin{aligned} U_1 &= \left(\frac{M}{L_2} + \frac{R_{12}}{R_{22}}\right) U_2 + \left(\frac{|R|}{R_{22}} + j\omega \frac{L_1 L_2 - M^2}{L_2}\right) J_1, \\ J_2 &= -\left(j\omega \frac{1}{L_2} + \frac{1}{R_{22}}\right) U_2 + \left(\frac{M}{L_2} - \frac{R_{12}}{R_{22}}\right) J_1, \end{aligned} \tag{28}$$

mit $|R| = R_{11} R_{22} + R_{12}^2$.

Aus diesem Gleichungssystem ist zu entnehmen, dass die Schaltung Isolatoreigenschaften bekommt, wenn die Bedingung

$$\frac{M}{L_2} = \pm \frac{R_{12}}{R_{22}} \tag{29}$$

erfüllt ist. Gilt das obere Vorzeichen, so erfolgt in Vorwärtsrichtung Energiedurchlass, während in der Rückwärtsrichtung Sperrung eintritt. Wird ein idealer Transformator benutzt, so verschwinden die imaginären Glieder der Koeffizienten und M/L_2 wird das Übersetzungsverhältnis des Transformators. Für maximale Energieübertragung muss dann als günstigster Abschlusswiderstand

$$R_{a\,\mathrm{opt}} = R_{22} \tag{30}$$

gewählt werden. Die Wirkleistungsübertragung ergibt sich in diesem Falle zu

$$\eta_{\max} = \frac{\epsilon^2}{1+\epsilon^2} \tag{31}$$

mit der schon früher verwendeten Abkürzung

$$\epsilon^2 = \frac{R_{12}^2}{R_{11} R_{22}}.$$

Auch wenn der Transformator entsprechend dem Ansatz in (27) und (28) nicht ideal, aber verlustlos ist, ergibt sich bei richtiger Anpassung der gleiche Wirkungsgrad (31). In diesem Falle ist der Anpassungswiderstand eine komplexe Grösse.

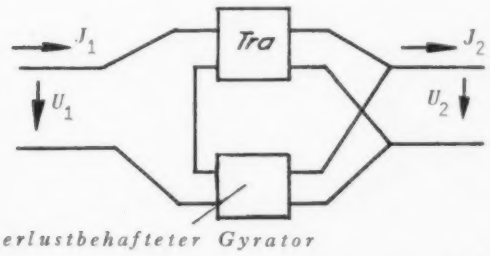


ABB. 13. Isolatorschaltung durch Reihenparallelschaltung von Transformator und Gyrator.

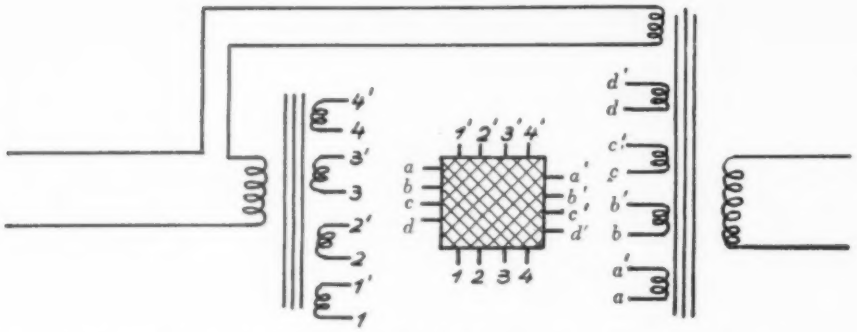


ABB. 14. Isolatorschaltung durch Reihenparallelschaltung von Transformator und transformatorgespeisten Vielelektroden-Hallgyrator.

Bei Verwendung von verlustbehafteten Transformatoren müssen Blindwiderstände zu der Schaltung gefügt werden. Die Bedingung für Einwegleitung (29) muss dann etwas erweitert werden und der Gesamtwirkungsgrad sinkt.

Aus (31) folgt für den 4-Elektrodengyrator, von dem in Abschnitt 2 gezeigt wurde, dass für ihn ϵ^2 mit wachsendem Hallwinkel gegen 1 strebt, bei Reihenparallelschaltung mit einem verlustlosen Transformator ein Grenzwert des Wirkungsgrades von 0,5. Dieser Grenzwert liegt somit doppelt so hoch wie der durch die übliche Widerstandsschaltung nach Abb. 12 bedingte Grenzwert.

Die Transformatorschaltung ist auch bei dem in Abschnitt 2 beschriebenen Vielelektroden-Gyrator anwendbar, und zwar ohne Benutzung eines neuen Transformatorkernes in der Schaltung nach Abb. 14. Bei Benutzung von (25) und (31) ergibt sich für diese Schaltung ein Grenzwert des Wirkungsgrades in Durchlassrichtung

$$\lim_{\mu B \to \infty} \eta_{\max} = \frac{nm}{nm+1} \tag{31}$$

mit den gleichen Bedeutungen für n und m wie in Abschnitt 2 a.

LITERATUR

1. W. P. MASON, W. H. HEWITT und R. F. WICK, *J. Appl. Phys.* **24**, 166 (1953).
2. R. F. WICK, *J. Appl. Phys.* **25**, 741 (1954).
3. F. KUHRT und W. HARTEL, *Arch. Elektrotech.* **43**, 1 (1957).
4. A. PERRIER, *Helv. phys. acta.* **27**, 207 (1954).
5. W. J. POPPELBAUM, *Helv. phys. acta.* **27**, 355 (1954).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 85–92. Printed in Great Britain

TRANSIENT CHARGE-CARRIER DIFFUSION IN HIGH-PURITY SILICON*

M. GREEN and I. N. GREENBERG

U.S. Army Signal Research and Development Laboratory, Fort Monmouth, New Jersey

(*Received* 13 *May* 1959; *in revised form* 11 *August* 1959)

Abstract—This paper presents the results of an experimental and theoretical study of the transient current resulting from the diffusion of a pulse of charge carriers photoinjected into high-resistivity silicon, i.e. ca. 1900 Ω-cm *p*-type and 590 Ω-cm *n*-type material. The charges diffuse to and are detected as a photovoltaic current by the metal-to-silicon contact which, in all cases, is found to be rectifying at room temperature.

Experimentally, it is found that the initial slope of the curve of α^2 (α is the distance from the point of injection to the junction) versus the time of occurrence of the maximum in the diffusion current averages about $6D$ for electrons in *p*-type silicon and $12D$ for holes in *n*-type silicon (D is the appropriate diffusion constant). The theory developed here predicts a slope of $6D$ for both electrons and holes. Discrepancies are attributed to (a) effect of light scattered from source to junction, (b) inaccuracies in the measurement of α and (c) errors in extrapolating to zero total resistance. Two peak photocurrents are distinguished: one due to diffusion of carriers to the junction, and one due to carriers generated at the junction by scattered light; as a consequence of the total circuit resistance, the first of these decreases more slowly than the second. The phenomenon is explained on the basis of a variational analysis of the total-current-density equation. An effective sample lifetime and an effective junction electric field for each external resistance are determined from the experimental data and the theoretical equations. Values of lifetime obtained in this way are 15–20 μsec for the *p*-type specimen and 30–40 μsec for the *n*-type (values on these same specimens obtained by the decay of the photoconductivity are larger by a factor of 2). The junction electric field values obtained when the total circuit resistances are extrapolated to zero range from tenths to several V/cm.

Résumé—On décrit dans cet article les résultats d'une étude expérimentale et théorique du courant transitoire résultant de la diffusion de porteurs injectés photoélectriquement dans du silicium à haute résistivité, ici 1900 Ω-cm de type *p* et 590 Ω-cm de type *n*. Les porteurs diffusent jusqu'à un contact métal-silicium où ils sont détectés sous forme d'un courant photovoltaïque. Ces contacts sont toujours redresseurs à la température ordinaire.

Expérimentalement on trouve qu'en portant le carré de la distance du point d'injection à la jonction en fonction de l'instant où se produit le maximum de courant de diffusion la pente de la courbe obtenue est en moyenne $6D$ pour des électrons dans du silicium de type *p* et $12D$ pour des trous dans du silicium de type *n*, D étant la constante de diffusion convenable. La théorie développée ci-dessous prévoit une pente $6D$ aussi bien pour les électrons que pour les trous. On peut attribuer ce désaccord à: (a) l'effet de la lumière diffusée depuis la source jusqu'à la jonction; (b) l'imprécision de la mesure de α; et (c) des erreurs dans l'extrapolation à une résistance totale nulle. On trouve deux maxima dans l'intensité du photocourant; l'un dû à la diffusion des porteurs jusqu'à la jonction, l'autre dû aux porteurs produits à la jonction par la lumière diffusée. Du fait de la résistance du circuit, le premier de ces maxima décroît plus lentement que le second. Ce phénomène est expliqué par une analyse variationnelle de l'équation donnant la densité de courant totale. On peut déduire des résultats expérimentaux et des équations théoriques une durée de vie effective et un champ électrique effectif dans la jonction pour chaque valeur de la résistance externe. Les valeurs de la durée de vie ainsi obtenues sont de 15–20 μsec pour le spéciment de type *p* et de 30–40 pour celui de type *n* (les valeurs obtenues par la décroissance de la photoconductivité sur ces mêmes specimens sont environ deux fois plus grandes). Les valeurs du champ électrique de la jonction obtenues en extrapolant à zéro la résistance totale du circuit varient de quelques dizièmes à quelques V/cm.

* A limited account of this work was presented at the annual meeting of the American Physical Society, New York.[1]

Zusammenfassung—Diese Arbeit stellt die Ergebnisse einer experimentellen und theoretischen Studie über den Durchgangsstrom dar, der sich aus der Diffusion eines Ladungsträger-Impulses ergibt, der in hochohmiges Silizium (1900 Ω cm *p*-dotiertes und 590 Ω cm *n*-dotiertes Material) photoinjiziert wird. Die Ladungen diffundieren zu einem Metall-Silizium-Kontakt, der bei Zimmertemperatur gleichrichtend ist, und werden dort als Photostrom nachgewiesen.

Experimentell wurde herausgefunden, dass die Anfangstangente von α^2 (α ist der Abstand vom Orte der Injektion zum Übergang) als Funktion der Zeit des Auftretens des Maximums im Diffusionsstrom durchschnittlich etwa 6 *D* für Elektronen in *p*-dotiertem Silizium und 12 *D* für Löcher in *n*-dotiertem Silizium beträgt (*D* ist die zugehörige Diffusionskonstante). Die hier entwickelte Theorie sagt einen Abfall um 6 *D* sowohl für Elektronen als auch für Löcher voraus. Die Diskrepanzen sind zuzuschreiben: (a) dem Effekt der Lichtstreuung von der Quelle zum Übergang, (b) den Ungenauigkeiten bei der Messung von α und (c) den Fehlern beim Extrapolieren auf den Totalwiderstand Null. Zwei Maximas des Photostroms sind zu unterscheiden: Das eine ist der Trägerdiffusion zum Übergang, das andere der Trägererzeugung durch Streulicht am Übergang zuzuschreiben; als Folge des totalen Kreiswiderstandes nimmt das erste langsamer ab als das zweite. Die Erscheinung wird auf der Basis einer Analyse der totalen Stromdichte-Gleichung erklärt. Eine effektive Lebensdauer und ein effektives elektrisches Feld im Übergang werden für jeden äusseren Widerstand aus den experimentellen Daten und den theoretischen Gleichungen bestimmt. Die Werte der Lebensdauer, die auf diese Weise erhalten wurden, betragen 15–20 μsec für das *p*-dotierte und 30–40 μsec für das *n*-dotierte Material (Werte für dieselben Materialien, die aus dem Abfall der Photoleitfähigkeit erhalten wurden, sind um den Faktor 2 grösser). Werte für die elektrische Feldstärke im Übergang betragen bei Extrapolation des Kreiswiderstandes auf Null einige Zehntel bis einige V/cm.

INTRODUCTION

THE diffusion of charge carriers accounts for many phenomena which are observed in semiconductors. Diffusion is the essential mechanism in the DEMBER[2] and the PME[3] effects. It plays an important role in *p–n* junction phenomena,[4] and in photovoltaic[5] effects. In certain thermoelectric and thermogalvanomagnetic effects the diffusion of the charge carriers contributes an appreciable fraction of the energy transport.[6] Obviously, these physical effects are also a means of studying diffusion. Generally in the investigation of such effects, however, more emphasis is placed on the effects than on the role that diffusion plays in the effect. In this investigation, the purpose is to study diffusion using the photovoltaic effect in high-purity silicon as a means of detecting and observing the phenomenon. This is probably the most direct method of observing and measuring the diffusion of charge carriers in a semiconductor.

EXPERIMENTAL

The photoinjection of electron–hole pairs in the vicinity of a junction contributes to the photocurrent due to the diffusion of the excess charge carriers to the junction.[5] Photoinjection is readily observable, in high purity silicon, following a short-duration light flash.[7] The apparatus is the same as that described in the reference. The silicon filter and the battery of the schematic diagram were not used in these experiments.

It is observed that as the distance between the point of injection and the junction is increased, there is an increase in the time at which the peak in the transient photocurrent occurs, Fig. 1. Also, a rapid attenuation of the peak and a spreading in time of the transient photocurrent are observed.

Photovoltages for the above light source[7] are shown for several metal contacts to *p*-type silicon, Table 1. For rhodium plated on *n*-type silicon, the photovoltage is about 100 mV. The polarity of the photovoltage from metal to silicon is positive for *p*-type and negative for *n*-type.

Table 1. *Photovoltaic out-put at metal-to-p-type-silicon contact* (*spark light source*)

Contact material	How applied	Photovoltage (mV)
Copper	Electroplated	4
Rhodium	Electroplated	15
Lead–Tin (60–40)	Ultrasonically soldered	25
Gold	Vacuum evaporated	30
Indium	Electroplated	100
Silver	Painted (dried at 250°C)	150

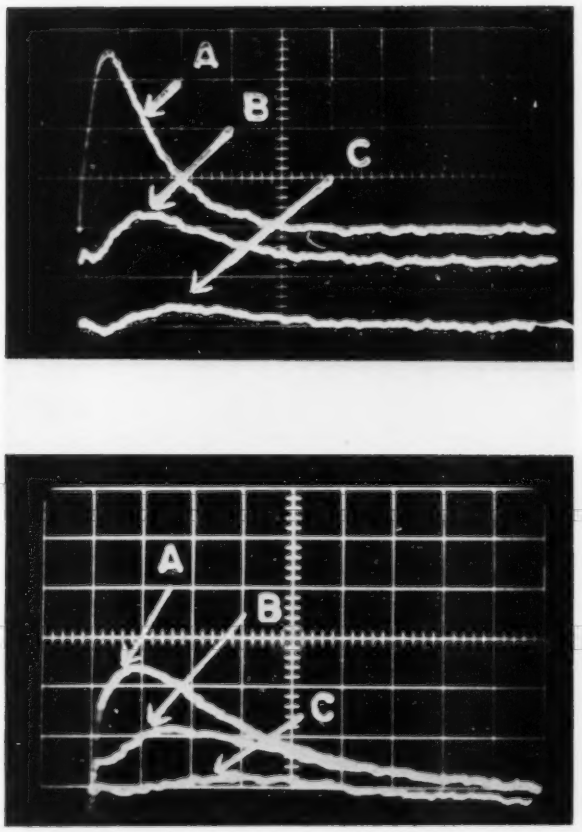


FIG. 1. Effect of distance of diffusion $\propto$ on the transient photocurrent of electrons in p-type silicon. Distance $\propto$ of source of injection from junction for upper oscillogram, in mm, A 2·29, B 2·69, C 3·30; for lower oscillogram, in mm, A 2·44, B 2·95, C 3·76. Oscillogram scales: Vertical $1{\cdot}30 \times 10^{-10}$ A/division, horizontal 50 μsec/division.

[facing p. 86

The results reported in this paper are confined to junctions formed (a) at a contact of silver painted on 1900 Ω-cm p-type silicon, then dried at 250°C, and (b) at a contact of rhodium electroplated on 590 Ω-cm n-type silicon.*

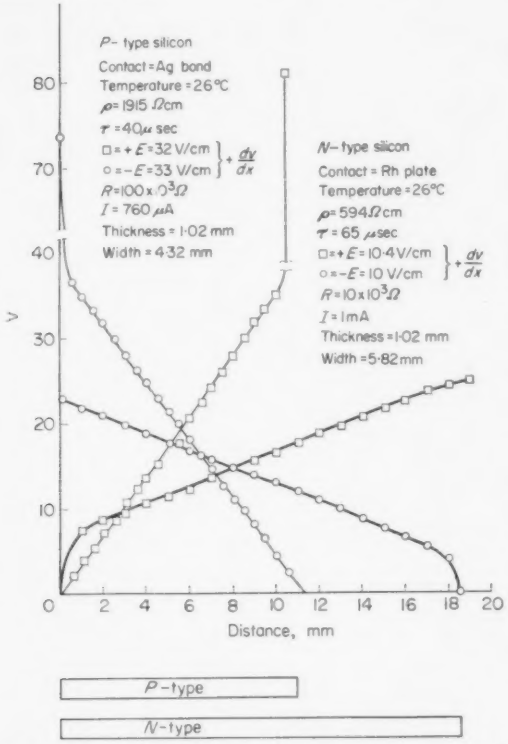


FIG. 2. Voltage-profile curves demonstrating, for high reverse current, junctions at the metal-to-silicon contacts, for both n- and p-type silicon.

Theoretical calculations for the transient photocurrent are given for one-dimensional geometry and for the assumptions shown at the top of Fig. 3 (except that diffusion is neglected inside the junction). Experimental observations of the effect of load resistance on the transient photocurrent are explained by an extension of results derived by VAN ROOSBROECK[8].

* In order to characterize these junctions more completely, voltage profiles are shown for fixed currents at room temperature in Fig. 2. The steep slope corresponds to the junction field in the blocking direction. The field E of equation (1) corresponds to no external applied-bias current.

THE TRANSIENT PHOTOVOLTAIC CURRENT

It is pointed out above that a photocurrent results from injection near a junction. Therefore, when the injection is transient, the photocurrent is transient. An approximate* expression for the photocurrent derived for one-dimensional geometry and a high barrier field at the junction is given by

$$I = [(Q_0/\sqrt{\pi})(\mu E\alpha)/(2\mu Et+\alpha)(1/\sqrt{Dt})\times \times \exp(-\alpha^2/4Dt-t/\tau)] \qquad (1)$$

for Q_0/e charges (hole–electron pairs) injected into the semiconductor uniformly, in an infinitely narrow plane, at a distance α from the junction. In equation (1), E is the effective average field strength of the junction, μ is the mobility of the minority charge carriers, D is the diffusion coefficient, t is the time, τ is sample lifetime, and Q_0 is the charge in coulombs due to photoinjected electron–hole pairs.

Practically, it is impossible to inject an infinitely narrow pulse of charge carriers into the semiconductor. Nevertheless, equation (1) still holds quite well for a finite width of injection, provided the width of the pulse of injected carriers is narrow compared with α. Equation (1) can be improved somewhat by taking into account the finite width W of the initially injected pulse, by substituting the quantity $t+\theta$ for t (except for the t in the second term in the exponent), where θ is given by†

$$\theta = W^2/4\pi D \qquad (2)$$

and D is the diffusion constant.

The peak of the transient current of equation (1) for a given α occurs at a time determined by

$$\alpha^2 = 4D[\tfrac{1}{2}+(2\mu Et)/(2\mu Et+\alpha)+(t/\tau)]t \qquad (3)$$

Equation (3) shows that the slope of the curve of α^2 versus t (time at which the current maxima occur) should increase with time. There are two ranges of

* Equation (1) is an approximate solution to the conditions shown at the top of Fig. 3. However, it is adequate for most of the conditions of these experiments. The exact solution is known in quadrature form, but it is most inconvenient for mathematical manipulation.

† For these experiments W is 0·17 mm, for which the values of θ are: for electrons, 0·6 μsec, and for holes, 2·5 μsec. These values of θ are negligible over the range of t for which measurements can be made.

time of interest in equation (3), namely

$$\tfrac{3}{2}\tau > t > \alpha/2\mu E \tag{4}$$

and

$$t > \tfrac{3}{2}\tau > \alpha/2\mu E \tag{5}$$

For the first range, α^2 is approximately equal to $6Dt$, and for the second range α is approximately

The peak current for the condition represented by equation (4) is given approximately by

$$I_{\text{peak}} \approx (Q_0/3{\cdot}3)[(3\mu ED)/(\mu E\alpha+3D)](1/\alpha)\times \times \exp(-\alpha^2/6D\tau) \tag{6}$$

and for the conditions imposed by equation (5), it

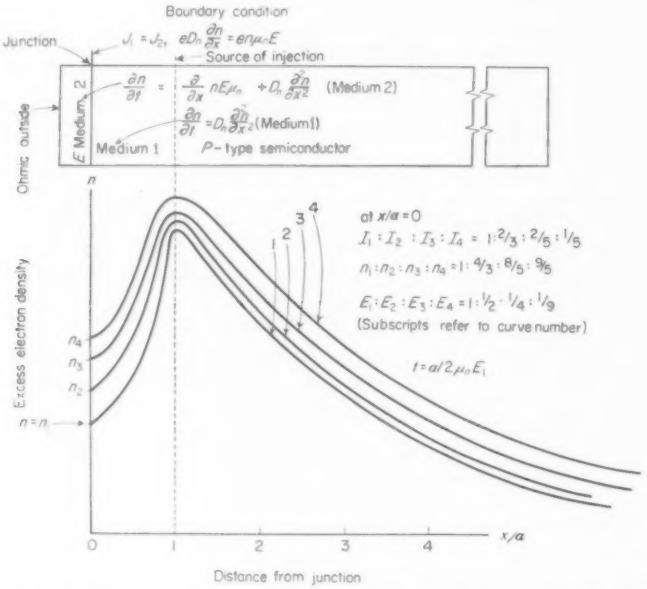


FIG. 3. Curves of excess carrier density (hole–electron pairs) versus distance from junction for various junction fields E at $x = 0$, and for a time after injection of $\alpha/2\mu_n E$ (the time for an electron to drift a distance $\alpha/2$ in a field E). Table indicates, at $x = 0$, relationship between currents, excess charge carriers, and junction field of the various curves. Diagram at top of figure indicates appropriate boundary conditions and differential equations for the two media on either side of junction—diffusion was neglected in medium 2 in the solution given by equation (1).

equal to $2\sqrt{(D/\tau)}t$; so that, for the first case, a plot of α^2 versus the time t at which the peak current in the experimental data occurs, should give a slope of $6D$, except for a short initial range (in practice never observed). In the second case, a linear plot of α versus t should give a slope of $2\sqrt{(D/\tau)}$ for $t > \frac{3}{2}\tau$. Experimental plots of α^2 versus t are shown in Fig. 4. The agreement with theory is satisfactory considering that the fiducial point ($\alpha = 0$) is very difficult to determine and for other reasons explained later.

is

$$I_{\text{peak}} \approx Q_0(0{\cdot}8D^{1/4})(\alpha^{-\frac{1}{4}})(\tau^{-5/4})\exp[-\alpha/\sqrt{(D\tau)}] \tag{7}$$

Only when the peak current occurs during a time short* compared with $\alpha/2\mu E$ does its value show a

* Experimentally, this is a difficult region to measure (a) because the injected pulse width W becomes comparable with the distance α to the junction, and (b) because the drift of charge carriers dominates overwhelmingly over their diffusion.

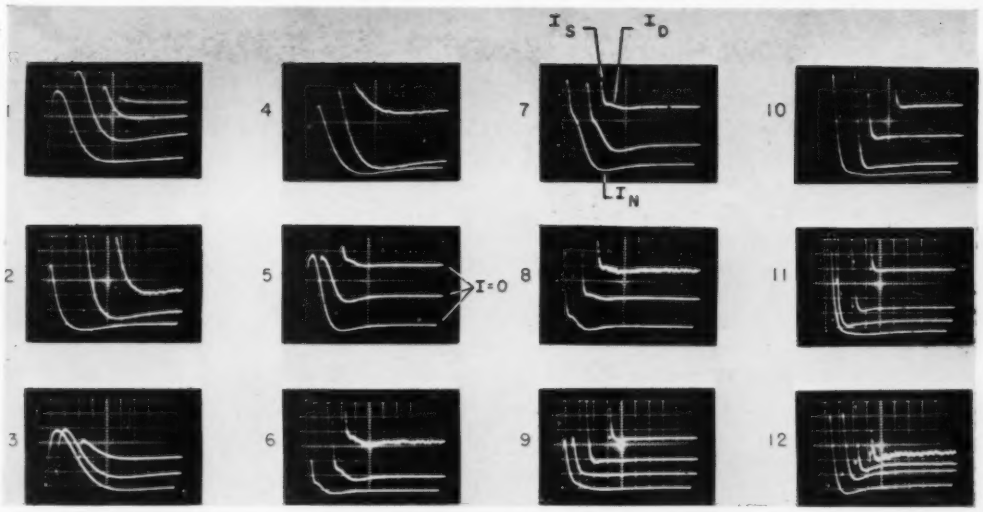


FIG. 5. Transient photocurrents resulting from diffusion of holes, in *n*-type silicon, to junction formed at metal contact. Series of oscillograms show effect of load resistance and distance of diffusion. Load resistances (curves from left to right) for oscillograms 1 and 2 are 100kΩ, 10kΩ, 1kΩ and 0·125kΩ, and for oscillogram 3, 100kΩ, 10kΩ and 1kΩ. Oscillograms 4–12 have load resistances (curves from left to right) of 1 MΩ, 100kΩ, 10kΩ and 1kΩ. Sample resistance approximately 10kΩ.

Oscillogram	1	2	3	4	5	6	7	8	9	10	11	12
Diffusion distance in mm	0·23	0·31	0·37	0·37	0·39	0·44	0·52	0·62	0·80	1·09	1·38	2·04

Horizontal scales: 20 μsec/division for oscillograms 1, 2, 3 and 12; 50 μsec/division for 4–11. Vertical scales in mV/division for curves, from left to right of each oscillogram

1—2·5, 2·5, 0·25, 0·05 2—1, 1, 0·25, 0·05 3—1, 1, 0·25 4 and 7—1, 1, 1, 0·1
5 and 6—1, 1, 0·5, 0·1 8—0·5, 0·5, 0·5, 0·1 9—0·5, 0·5, 0·25, 0·05 10—0·25, 0·25, 0·25, 0·05
11—0·25, 0·25, 0·1, 0·05 12—0·1, 0·1, 0·1, 0·05

The first current peak I_s (oscillogram 7) is that resulting from light scattered to junction, the second current peak I_D is the one resulting from diffusion. The horizontal tail (oscillogram 5) of each current transient corresponds to zero current level for that trace.

[facing p. 88

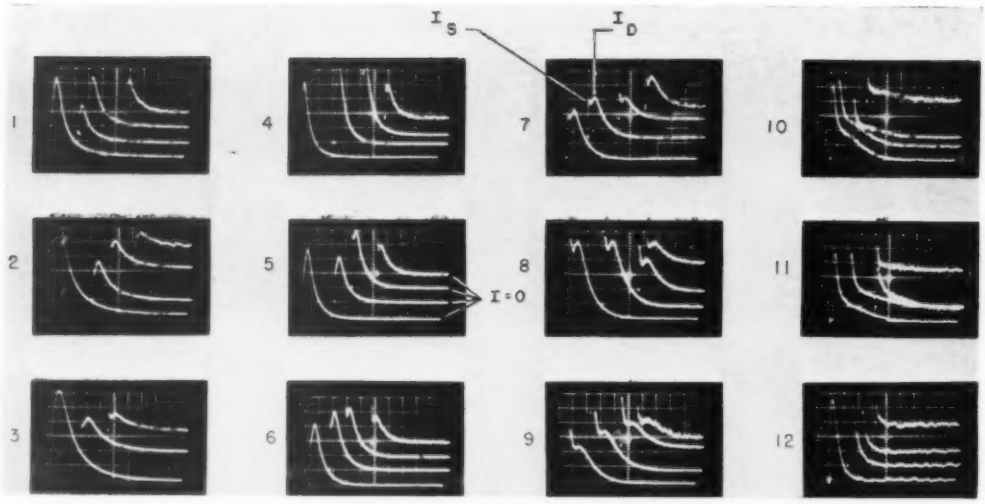


FIG. 6. Transient photocurrents resulting from diffusion of electrons, in *p*-type silicon, to junction formed at metal contact. Series of oscillograms show effect of load resistance and distance of diffusion. Oscillograms 1, 9, 10, 11 and 12 have load resistances (curves from left to right) of 1 MΩ, 100kΩ, 10kΩ and 1kΩ. Oscillograms 2, 4, 6 and 8 have load resistances (curves from left to right) of 10kΩ, 1kΩ and 0·125kΩ. Oscillograms 3, 5 and 7 have load resistances (curves from left to right) of 1 MΩ, 100kΩ and 10kΩ. Sample resistance, 120kΩ.

Oscillogram	1	2	3	4	5	6	7	8	9	10	11	12
Diffusion distance in mm	0	0	0·34	0·34	0·60	0·60	0·76	0·76	1·04	1·49	1·78	3·01

Horizontal scales: 20 μsec/division for oscillograms 1–4, 50 μsec/division for 5–12. Vertical scales in mV/division for curves, from left to right, of each oscillogram:

1 and 3—10 2—10, 0·5, 0·1 4—2·5, 0·5, 0·1
5—5 6—2·5, 0·25, 0·05 7—2·5
8—1, 0·25, 0·05 9—2·5, 2·5, 1, 0·25 10—1, 1, 0·5, 0·25
11—1, 1, 1, 0·25 12—0·5, 0·5, 0·25, 0·05

The first current peak I_s (oscillogram 7) is that resulting from light scattered to junction, the second peak I_D is the one resulting from diffusion. I_n, oscillogram 7, is a negative component of the current not observed in *n*-type material, see text. The horizontal tail (oscillogram 5) of each current transient corresponds to zero current level for that trace.

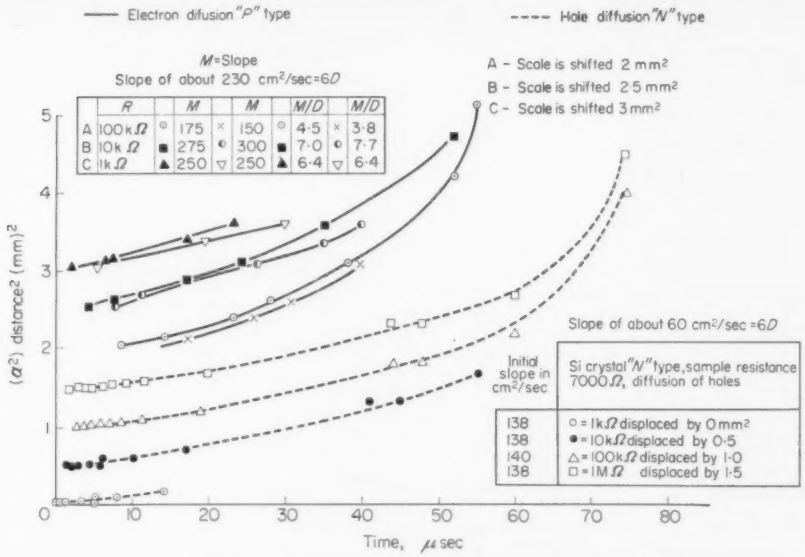

Electron difusion "P" type
Hole diffusion "N" type
M=Slope
Slope of about 230 cm²/sec=6D
A - Scale is shifted 2 mm²
B - Scale is shifted 2·5 mm²
C - Scale is shifted 3 mm²
R | M | M | M/D | M/D
A 100kΩ | 175 | 150 | 4·5 | 3·8
B 10kΩ | 275 | 300 | 7·0 | 7·7
C 1kΩ | 250 | 250 | 6·4 | 6·4
(α²) distance² (mm)²
Slope of about 60 cm²/sec =6D
Initial slope in cm²/sec
Si crystal "N" type, sample resistance 7000Ω, diffusion of holes
138 ○ = 1kΩ displaced by 0 mm²
138 ● = 10kΩ displaced by 0·5
140 △ = 100kΩ displaced by 1·0
138 □ = 1MΩ displaced by 1·5
Time, μsec

FIG. 4. Square of distance, α^2, from source of injection to junction (diffusion distance) versus time t at which current maxima occur. Theory predicts an initial slope of $6D$ for both p- and n-type silicon, viz. about 230 cm²/sec for p-type and 60 cm²/sec for n-type. Slopes of experimental curves increase with time, in agreement with equation (3).

strong dependence on E, namely

$$I_{\text{peak}} \approx Q_0(\mu_n E/2\alpha)\exp(-\alpha^2/2D\tau). \qquad (8)$$

The three equations (6), (7) and (8) indicate that the peak current falls off rapidly with distance. This is confirmed qualitatively by the oscillograms of Fig. 1.

EFFECT OF CIRCUIT RESISTANCE ON PHOTOVOLTAIC CURRENT

The circuit resistance of a solar cell is known to affect the current in a non-linear[5] fashion. Similar effects are observed with these rectifying contacts. The transient photovoltaic current is changed in a very interesting manner by variation of the series resistance as shown in Figs. 5 and 6. This effect is explained subsequently. It is also observed that the total circuit resistance has a different effect on the photocurrent resulting from light scattered to the junction which appears on the time scale as the first peak, than it does on the photocurrent arising from the diffusion of the charges to the junction which appears later as a second peak. In particular, for large series resistance and injection close to the junction, it is observed that the peak for the diffusion photocurrent exceeds that of the photocurrent resulting from scattering, while for sufficiently low external series resistance, the photocurrent from the scattered light predominates as shown in oscillograms 5 and 8 of Fig. 5 and 5 of Fig. 6. The time at which the second peak appears is also shortened as the series resistance is reduced. The distributed capacitance of the circuit contributes but very slightly to the shape of the current transients, as evidenced by the fact that at $\alpha = 0{\cdot}3$, Fig. 5, where the photoinjection takes place practically at the junction, the much sharper current wave is not perceptibly affected by the change in the external load (note that this oscillogram has the fastest time base).

The observed effect of load resistance is explained by assuming that the effective barrier field E in equation (1) depends on the integral of the current, that is on the charge which has flowed into the junction for the time t. This, in turn, causes a non-linear dependence of the current on the series resistance. A more comprehensive explanation will be given later in the discussion of this effect and of the effective junction field.

An effective sample lifetime τ and an effective

junction field E are determined for each external resistance R_L, by the composite of data represented by the curves of Fig. 7. In these graphs, the logarithm of the product of the peak voltage across R_L and the distance of injection α are plotted against time. These curves correspond essentially to equation (6). They differ by the factor R_L. For small α, equation (6) indicates that negative slope of the curves should be τ^{-1} (recall that in this range $\alpha^2/6D = t$). The values thus obtained, taking the spread of the data into account, are 15–20 μsec for the p-type and 30–40 μsec for the n-type sample. For comparison, in the corresponding order, lifetime values of 40 μsec and 70 μsec are measured by the decay of the photoconductivity. It is felt that the discrepancy results from the fact that the current arising from the light scattered* to the junction causes an error in the measurement of the current arising from charge diffusion. In particular, the plotted current values corresponding to short diffusion times are too large, hence indicating an erroneously large slope and consequently a lifetime value which is too small.

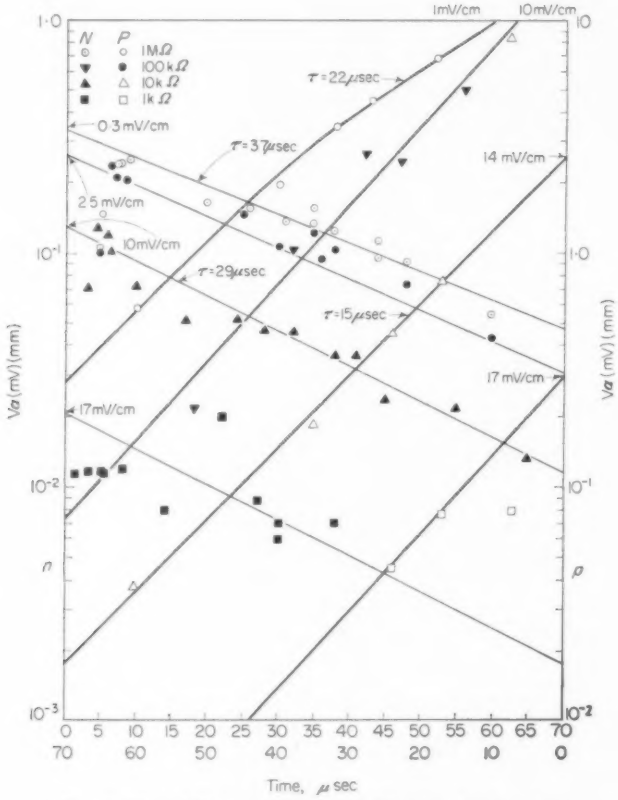


FIG. 7. Plot of $V\alpha$ versus t (voltage V across load resistor R_L times α versus t) for n- and p-type silicon. The intercepts on the $V\alpha$ co-ordinate are equal to $R_L Q_0 \mu E/3{\cdot}3$; the corresponding junction field E is indicated at the intercept. Theoretically, negative slopes of the curves are approximately equal to $2{\cdot}30\ \tau^{-1}$, and hence permit determination of sample lifetime (see text for results).

* The scattered light from a slit opening is difficult to eliminate. A new design is now being fabricated in an effort to reduce it.

According to equation (6) the intercepts of the curves of Fig. 7 with the ordinate axis should be equal to $R_L Q_0 \mu E/3{\cdot}3$ (since $V_{\text{peak}} = I_{\text{peak}} \times R_L$). Hence, if μ and Q_0 are known,* the junction field E can be determined. The values of E, so determined, are indicated at each intercept. It is observed that they decrease with increasing external series resistance R_L. This is the result of polarization of the junction field by the charge carriers which diffuse into it.† A better understanding of the effect of external series circuit resistance on the photocurrent and on the effective junction field is obtained from the following approach.‡

VAN ROOSBROECK[8] has shown that the total current density J is solenoidal, so for the one-dimensional case

$$J(t) = J_p(x, t) + J_n(x, t) + \partial/\partial t(\epsilon E/4\pi) \qquad (9)$$

where

$$J_p + J_n = e[\mu_p(nb+p)E + D_p d/dx(nb-p)] \qquad (10)$$

and where J_p, J_n and $\partial/\partial t(\epsilon E/4\pi)$ are hole current, electron current and displacement current, respectively. The total current $J(t)$ is time-dependent but not space-dependent. Operating variationally on equation (9) (resulting from a variation of charge carriers due to injection), one finds

$$R_c \delta I = (b+1) \int_0^l \frac{\delta n E\, dx}{nb+p} + \frac{kT}{e}(b-1) \int_0^l \frac{1}{nb+p} \times$$

$$\times \frac{d\delta n dx}{dx} + R_c \frac{\partial}{\partial t} \int_0^l \delta(p-n) A\, dx \qquad (11)$$

where R_c is the series circuit resistance, δI is the current change, b is the mobility ratio, δn is the change in carrier concentration (hole–electron pairs), kT/e is the Boltzmann potential, A is the cross-sectional area of the sample, l is the sample length, p is the hole concentration, n is the electron concentration and E is the electric field. The last term of equation (11) is due to the displacement current and contributes only for a period of time of the order of the dielectric relaxation time. It may be neglected, except where fast transients or very high resistivity materials are concerned. The second term of equation (11) is concerned with the junction field, and the third term, with the Dember field. These latter two terms are of opposite sign for a junction formed by a metal contact to p-type silicon (electron diffusion) and of the same sign for a junction formed by a metal-to-n-type-silicon contact (hole diffusion). In general, the Dember field contributes much less to the photovoltage than the junction field, so that the second term is the dominating one, and is exact if E is considered to be the existing field corresponding to δI. E varies in an inverse manner with δn because of the resulting polarization field. However, the second term of equation (11) increases with increasing δn despite the decrease in E (although for very large δn it will undoubtedly approach an asymptotic value). As E decreases at the junction, its value is reflected in equation (1) so that the diffusion current into the junction decreases, which, in turn, causes δn at the junction to increase, since this is required by the boundary conditions of the solution of equation (1), Fig. 3. Therefore, since δI of equation (11) is equal to I of equation (1), the potential, as given by the right side of equation (11), must increase with total series circuit resistance R_c. Hence for equal conditions of injection near to the junction, the product δI times R_c should

* The Q_0 values, determined from the initial photoconductivity current resulting from injection into the samples with d.c. current biases, are 10^{-9} and 4×10^{-9} C for the n- and p-type samples respectively. Mobility values of 400 cm²/V-sec for n-type and 1500 cm²/V-sec for p-type silicon had been previously measured by the "drift" [7] method.

† Plots of log E versus R_c, not shown, where R_c is the total series circuit resistance, indicate that the short circuit junction fields should be of the order of tenths of a V/cm to 1 V/cm for both the p- and the n-type specimens. Because the sample resistances are high, the extrapolations are not accurate.

‡ This derivation does not mention specifically the manner in which the junction capacitance may contribute to the observed effect of total circuit resistance on the photocurrents generated by light scattered directly to the junction as compared with that caused by the diffusion of charges to the junction. It should be noted, however, that δn contains time as a factor, so that the shape of the respective current pulses appears distorted by the combination of junction capacitance and series circuit resistance. The time constant associated with this impedance combination can conceivably show greater attenuation for the higher frequencies and hence for narrower pulses such as the scatter-generated current pulse as opposed to the current pulse generated by carrier diffusion to the junction.

increase with increasing total series circuit resistance R_c.

The last statement of the previous paragraph should be true for long-lifetime material. However, if the charge carriers are not pulled out sufficiently fast, they will recombine before they reach the junction, so that in going from high-lifetime material to low-lifetime material, the above situation should tend to reverse. Experimental measurements and further theoretical work are now in progress.

In passing, the authors wish to point out the negative *component* of the transient photocurrent which appears after the positive one, in *p*-type, but *not* in *n*-type silicon (Fig. 6 oscillograms 4, 5, 7 and 12). Such a negative transient has also been observed in *p*-type germanium, but *not* in *n*-type (see Ref. 7, footnote** p. 1478). An explanation has been given by GREEN in terms of the difference in the diffusion constants of the two carriers. A paper on this subject is in preparation.

Acknowledgements—The contribution of Mr. J. MANDELKORN, U.S. Army Signal Research and Development Laboratory, to the early work on this problem is gratefully acknowledged. We wish to thank Dr. W. KAISER of Bell Telephone Laboratories, and the Semiconductor Division of Merck and Company for providing the high-purity silicon samples used in this work.

REFERENCES

1. M. GREEN, I. N. GREENBERG and J. MANDELKORN, *Bull. Amer. Phys. Soc.* (2) **4,** 27 (1959).
2. H. DEMBER, *Phys. Z.* **32,** 554, 856 (1931); **33,** 207 (1932).
3. W. VAN ROOSBROECK, *Phys. Rev.* **101,** 1713 (1956); this reference has a large bibliography on the PME effect. W. GARTNER, *Ibid.* **105,** 823 (1957).
4. W. SHOCKLEY, *Electrons and Holes in Semiconductors* Chap. 4, Sections 2 and 12. Van Nostrand, New York (1950).
5. R. L. CUMMEROW, *Phys. Rev.* **95,** 16 (1954); W. G. PFANN and W. VAN ROOSBROECK, *J. Appl. Phys.* **25,** 1422 (1954); M. B. PRINCE, *Ibid.* **26,** 534 (1955). Additional references are cited in these articles.
6. G. E. PIKUS, *J. tech. Phys., Moscow* **26,** 22 (1956); O. LINDBERG, *Proc. I.R.E.* **40,** 1414 (1952), bibliography referenced therein; H. METTE, W. GARTNER and C. LOSCOE, *Phys. Rev.* **115,** 537 (1959). M. GREEN, *The Thermoelectromotive Force.* In preparation.
7. M. GREEN, *J. Appl. Phys.* **28,** 1473 (1957).
8. W. VAN ROOSBROECK, *Bell Syst. Tech. J.* **29,** 560 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1. p. 93. Printed in Great Britain

NOTES

Preferential Plating on Germanium

(*Received* 23 *September* 1959)

CAMP[1] has used a high-voltage pulsed-plating technique to deposit copper on the surface of freshly-etched germanium. The plating appears as an array of lines which were shown by voltage-probe measurements to lie on low-resistivity striations running through the crystal. Attempts made in these laboratories to reproduce his results were successful with *p*-type germanium, but failed altogether with *n*-type, giving merely a uniform deposit.

n-Type germanium

This has been overcome by plating *n*-type germanium (uniformly) then reversing the cell polarity and etching for a similar length of time, when the characteristic fine lines appear and gradually widen and increase in number.

The experimental details of the process were as follows:

Ohmic contacts were made to specimens by grinding, copper-plating and soldering. The surfaces under study were ground, chemically polished in a vigorous etch such as CP8, washed in de-ionized water, and pulse-plated immediately. Generally a suitable electrolyte was an acidic saturated solution of copper sulphate. Pulsing was effected by discharge of a capacitor, 2–8μF/cm^2 of plated area, at 1/4–1 pulse/sec, charged to 170–350 V (the higher values are required for high-resistivity material); the plating time was several minutes. The total charge passed during plating, and subsequent etching of *n*-type Ge, was of the order of 0·05 C/cm^2.

p-Type germanium

An infra-red absorption apparatus[2] based on the method of HARRICK[3] was used to verify that copper is deposited on low-resistivity regions. A crystal of *p*-type Ge, with a mean resistivity of 0·015 Ω-cm, was pulled at about 0·02 mm/rev, and one part of it (A) pulse-plated, while another flat piece (B), originally adjacent to the plated surface, mechanically polished. Light of about 2·2μ wavelength was then focused to a small spot on the surface of the specimen (yielding a linear resolution of 40μ) and its optical transmission measured along a line parallel to the axis of growth. Since the absorption coefficient at this wavelength is proportional to the free-carrier concentration, high transmission corresponds to high resistivity. The results for the two specimens are compared in Fig. 1, and the correlation is seen to be good. Close inspection of the surface of A when finely plated, compared with the absorption of B, indicates that copper will plate under suitable conditions provided $\delta\rho/\rho \sim 0{\cdot}1$–$0{\cdot}2$. If $\delta\rho/\rho > \sim 0{\cdot}4$, dense lines are plated. The mean spacing of the lines was equal to the length of crystal grown during 1 rev, and on tracing them round the bulk crystal they were found to form a helix. Thus it has been verified that copper plates preferentially on low-resistivity striations in the bulk, which appear to be associated with fluctuations in growth rate during pulling of the crystal.

F. E. ROBERTS

Mullard Research Laboratories
Cross Oak Lane
Salfords, Nr. Redhill
Surrey

Acknowledgements—The author wishes to thank Drs. M. D. STURGE and P. D. SOUTHGATE for valuable advice; Rev. A. SPRAY, Messrs. J. DALE and J. SEWELL for help with photography and polishing; and Mr. A. SMITH for the supply of crystals.

References

1. P. R. CAMP, *J. Appl. Phys.* **25,** 459 (1954).
2. F. E. ROBERTS, To be published.
3. N. J. HARRICK, *Phys. Rev.* **101,** 491 (1956).

Solid-State Electronics Pergamon Press 1960. Vol. 1. p. 94. Printed in Great Britain

BOOK REVIEW

G. TROUP: **Masers.** Methuen, London; John Wiley, New York, 1959. 168 pp., 13*s*. 6*d*.

THE presentation of a subject which embraces quantum mechanics, microwave spectroscopy and paramagnetic resonance in a pocket-size edition is a formidable task. The author has given a brief account of the maser art up to the beginning of 1958, with occasional reference to later work. Formulae are frequently quoted from the original papers without detailed derivations. It is difficult to see how this could have been avoided in a book of this size. However, the reader who desires more than a superficial understanding of the subject will often have to consult the original papers.

Chap. 2 describes the theory of stimulated and spontaneous emission. Chap. 3 is concerned both with the various factors which affect the line-shape of a transition and with the saturation of the transition which occurs when the stimulating field is large. Saturation plays a dominant role in the activation of the paramagnetic ions in the three-level maser, which is the centre of interest at the present time. This reviewer would like to have seen a detailed treatment of the effect of saturation on a multi-level system. The author has confined his attention primarily to the two-level case.

A particular transition can be rendered emissive in various ways: electrostatic focussing, microwave pumping, adiabatic rapid passage, pulse inversion and optical pumping. These activation methods are all described very briefly in Chap. 4 which is only ten pages long. Chap. 5 is perhaps the most useful in the book. It is devoted to a calculation of the gain, bandwidth and noise figure of both travelling-wave and cavity masers. The author uses an unconventional definition of the noise figure. He assumes that the temperature of the signal generator is equal to the ambient temperature of the device. In the conventional definition the temperature of the signal generator is 290°K. However, once this departure from convention is appreciated, no confusion is likely to arise. The theory of the ammonia maser oscillator is described in Chap. 6 and the experimental work on both ammonia and solid-state masers up to the beginning of 1958 is reviewed in Chap. 7.

This book is probably the first to be published on the subject of masers. For the new worker in the field a review of the literature is invaluable. The progress in the field of solid-state masers in the last two years has been tremendous. Several topics of current interest are therefore omitted from the book altogether or are only touched on lightly. This is hardly a criticism, it is the fate of any book which deals with a rapidly expanding subject.

P. N. BUTCHER

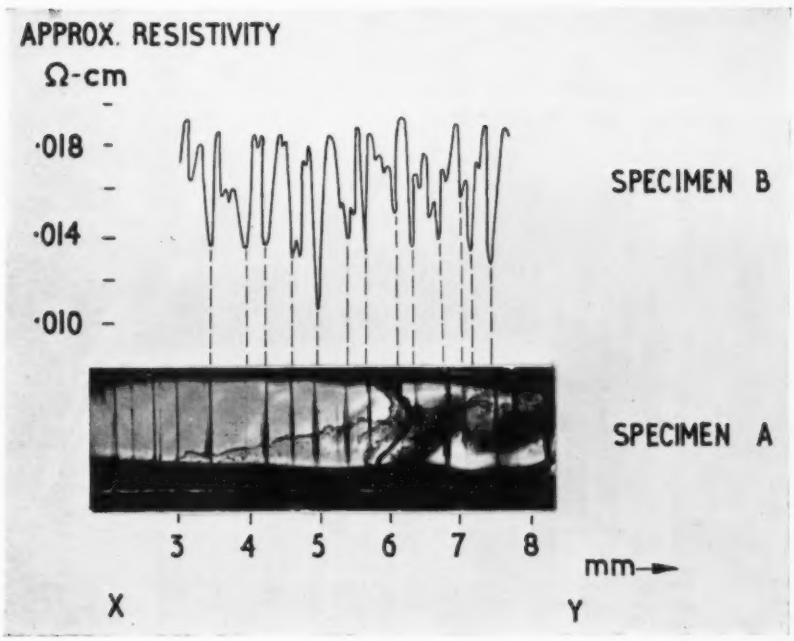


FIG. 1. Comparison between the resistivity of B by infra-red transmission, and plated lines on A. [Roberts, p. 93]

[*facing p.* 94

Notice of Conference

Conference on Solid-State Microwave Amplifiers

The Institute of Physics announces that its Electronics Group, in conjunction with the Radio Spectroscopy Group, is arranging a Conference on *Solid-State Microwave Amplifiers* to take place at the University of Nottingham from 6–8 April, 1960.

The subjects to be covered by the Conference include:

(a) Masers

(b) Parametric amplifiers

(c) Related resonance studies

Accommodation will be available in one of the University Halls of Residence.

Further particulars and application forms are available from The Secretary, The Institute of Physics, 47 Belgrave Square, London S.W.1.

Papers to be published in future issues

F. A. CUNNELL, J. T. EDMUND and W. R. HARDING: Technology of gallium arsenide

J. M. LAVINE: The behaviour of p- and n-doped contacts in a space-charge depletion region

R. SAINTESPRIT: Sur la diffusion des elements des colonnes III et V dans le silicium

A. J. STRAUSS, T. C. HARMON, E. B. OWENS and M. C. FINN: Indium as a source of impurities in indium arsenide

J. E. PARROTT: The interpretation of the stationary and transient behaviour of refrigerating thermocouples

R. L. WILLIAMS: Radiation limited lead-sulphide cells

R. L. WILLIAMS: Minority-carrier effects in chemically deposited PbS photoconductive films

MOTOSUKE MONAKATA: Electrical conductivity of the high vanadium-phosphate glass

BLANK
PAGE

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 97–106. Printed in Great Britain

TECHNOLOGY OF GALLIUM ARSENIDE

F. A. CUNNELL, J. T. EDMOND and W. R. HARDING

Services Electronics Research Laboratory, Baldock, Herts.

(*Received* 21 *September* 1959)

Abstract—The steps in the preparation of the compound semiconductor gallium arsenide are described, from the treatment of the component elements to the zone purification and production of single crystals of the compound. The efficiency of zone refining and the influence of some impurities on the electrical properties of the material are discussed. The purest material prepared by the authors has an electron mobility of 6700 $cm^2 V^{-1} sec^{-1}$ at room temperature. Techniques for making simple electrical measurements and etching-procedures are recommended. The production of p–n junctions by diffusion and some effects of heat treatment on the electrical properties are also included.

Résumé—On décrit les étapes dans la préparation du composé semiconducteur AsGa, depuis le traitement des éléments composants, jusqu'à la purification par zone fondue et la production des monocristaux du composé. On discute l'efficacité du raffinage par zone et l'influence de certaines impuretés sur les propriétés électriques du matériau. Le matériel le plus pur préparé par les auteurs a une mobilité de 6700 $cm^2 V^{-1} sec^{-1}$ à la température ordinaire. On recommande des méthodes de traitements et de mesures électriques simples. Dans l'article est inclus aussi la préparation de jonctions p–n par diffusion et certains effets de traitements thermiques sur les propriétés électriques.

Zusammenfassung—Es werden die einzelnen Stufen der Präparation der halbleitenden Verbindung Galliumarsenid beschrieben, und zwar ausgehend von der Behandlung der beiden Komponenten bis zur Zonenreinigung und Herstellung von Einkristallen der Verbindung. Die Wirksamkeit der Zonenreinigung und der Einfluss einiger Verunreinigungen auf die elektrischen Eigenschaften des Materials werden besprochen. Das von den Autoren hergestellte reinste Material hat eine Elektronenbeweglichkeit von 6700 $cm^2 V^{-1} sec^{-1}$ bei Zimmertemperatur. Hinweise für einfache elektrische Messungen und Ätzverfahren werden gegeben. Ausserdem werden die Herstellung von p–n-Übergängen durch Diffusion sowie einige Auswirkungen des Temperprozesses auf die elektrischen Eigenschaften behandelt.

1. INTRODUCTION

SINCE 1952 the group III–group V compounds have attracted an increasing amount of attention. Two of these materials, GaAs and InP, have energy gaps of 1·37 eV and 1·27 eV respectively and in view of their high electron mobilities they are considered as possible competitors of silicon for certain applications. This paper concerns GaAs but many of the technological problems which are discussed arise in the processing of other compounds possessing a volatile constituent e.g. InAs, InP and GaP. There have been a number of previous publications dealing with various aspects of GaAs.[1–4] In view of the growing interest in this compound a general account of techniques required for preparation, purification, crystal growing and other problems seems timely. The account is divided into two sections. The first deals with the purification of the elements and the preparation, zone purification and growth of crystals of the compound. In the second section there is a discussion of the effect on the electrical properties of GaAs of doping with certain impurity elements and of the problems which arise in the course of making measurements on the material. This includes etching and the practical details of making low-resistance ohmic contacts to n- and p-type material. In addition the effects of heat treatment on the compound and the production of p–n junctions by impurity diffusion are described.

2. PREPARATION OF GALLIUM ARSENIDE

(a) *Purification of the elements*

The application of zone refining to the compound GaAs has met with only limited success and will be discussed in detail below. Consequently it

is desirable to obtain starting elements of the highest available purity. In general, gallium and arsenic containing 3–30 p.p.m. total impurity, as indicated by spectrographic analysis, are readily available commercially. A problem arising in any purification work at the level of a few parts per million is the assessment of the impurity which remains. Spectrographic techniques are of limited application and in general the efficiency of the treatment has been judged by preparing batches of GaAs and measuring Hall constant (R) and conductivity (σ).[4]

Methods for the purification of gallium have been described in the literature.[5,6] Although a number of attempts have been made by the present authors to improve the gallium, they find that gallium arsenide prepared from the best currently available metal without further treatment is at least as good as any other material they have measured.

Arsenic is available in powder or metallic form, the powder form especially contains a quantity of arsenic trioxide, which can be removed by heating to 250°C in a continuously pumped vacuum system. The arsenic can then be sublimed at 500°C in vacuum and deposited in the metallic form at a lower temperature. Non-volatile residues such as carbon and traces of iron are left behind. A number of ingots of gallium arsenide were prepared with arsenic treated in this way. These showed an improvement on zone refining, indicating that some n-type impurities were segregating[2], but not all. Sulphur was believed to be a troublesome impurity in InAs prepared in this laboratory and elsewhere[7, 8] and to be present originally in the arsenic; equally it was thought that sulphur was likely to be present in GaAs and the arsenic was treated to remove this impurity. The method used is described by HARMAN *et al.*[8]; the arsenic is dissolved in high-purity molten lead at 650–700°C and sublimed from the solution. Because of the affinity of sulphur for lead, the sublimate is relatively free of sulphur. This process is followed by a further sublimation in vacuum. GaAs prepared from arsenic which has had the above treatment has proved invariably to be purer than the compound made from untreated arsenic.

The segregation coefficient of some impurities in GaAs has been measured by WHELAN *et al.*[9] They confirm that sulphur and also selenium and zinc show poor segregation.

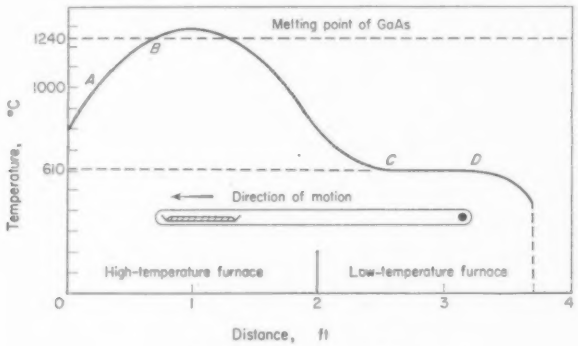


FIG. 1. Preparation of GaAs.

(b) *Synthesis of gallium arsenide*

The pressure–temperature composition relationship for the system Ga–As has been investigated by BOOMGAARD and SCHOL[3]. They find that the dissociation pressure of the compound GaAs is 0·9 atm at the melting point (1237°C). The vapour pressure of arsenic rises very rapidly with temperature from 1 atm at about 610°C to 37 atm at 818°C, the critical point. As a result the direct synthesis of GaAs in a uniform temperature enclosure must be attempted with caution to avoid dangerously high arsenic pressures. It is possible, however, to prepare GaAs in a quartz tube in this way, provided the temprature is raised slowly so that near-equilibrium conditions are maintained. Since GaAs tends to "key" to quartz, it has been

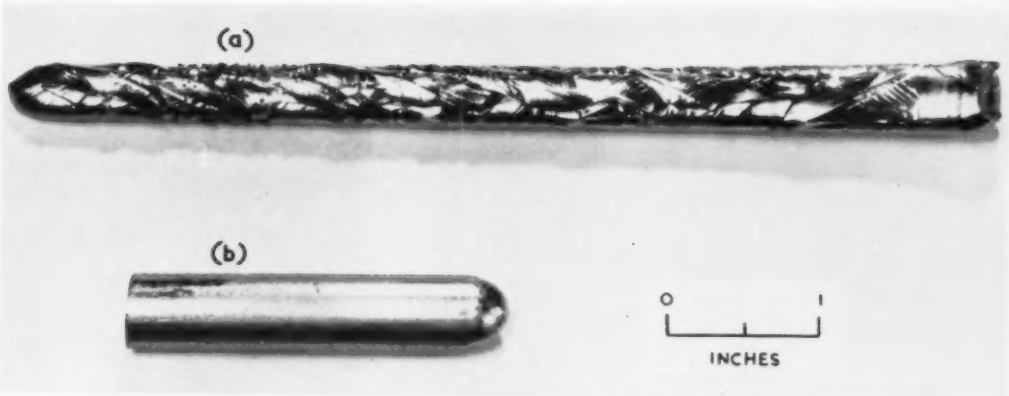


FIG. 2. Gallium arsenide ingots:
(a) polycrystalline
(b) single crystal.

[*facing p.* 98

BLANK
PAGE

found advisable to use a quartz boat or preferably an unsealed tube to contain the material. In this way any risk of the vacuum wall being broken on cooling is avoided. The logical development of the above method of preparation which is now used is to employ a composite furnace system to provide the temperature profile shown in Fig. 1. At one end a constant temperature of 610°C is maintained over an appreciable length *CD* and at the other end a temperature of about 1250°C. The inner quartz tube initially containing only the gallium is located at the high-temperature end, and the arsenic pressure in the system determined by the temperature of the colder end. Sufficient excess arsenic is added to keep about 1 atm of pressure in the enclosure even after the stoichiometric compound has been formed. By moving the outer quartz tube relative to the temperature gradient *AB*, GaAs can be crystallised in a controllable manner from one end of the boat. It is essential to provide a uniform low-temperature region of sufficient extent to ensure that the arsenic reservoir remains at a constant temperature throughout the solidification. Variations in the temperature of the furnace surrounding the arsenic reservoir should not exceed ±2°C. This method yields ingots containing large crystallites, Fig. 2(a).

(c) *Zone purification of the compound*

The problem of the zone purification of a compound in which one element is appreciably volatile has been discussed elsewhere.[4, 10] In all experiments in which GaAs is melted, whether it be zone refining, crystal growing or even a simple doping process, provision must be made for the required ambient pressure of arsenic. This involves the design of a suitable furnace. The production of a molten zone can be achieved either by radiation heating or by radio-frequency induction heating. In the first method accurate temperature control is necessary since the melting point of GaAs is close to the temperature at which quartz is easily deformed. With induction heating the hottest point is within the quartz tube, the walls of which remain at a much lower temperature, and this method is preferred. The material, together with excess arsenic, is placed in an unsealed quartz tube inside an evacuated sealed tube. This in turn is situated in a third quartz tube, much longer, which is open at both ends and which can be moved through the coil and its two associated furnaces, placed one on either side of the coil (Fig. 3). The outermost tube serves only as a convenient,

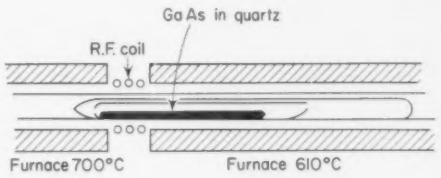


FIG. 3. Apparatus for zone refining.

transparent support for its contents. It is important to keep the gap between the furnaces to a minimum in order to avoid a cold spot in the temperature distribution. This requirement implies a coil with few turns. Radio-frequency generators operating at 400 kc/s and 5 Mc/s have been used. It is possible to produce a zone in relatively impure material using the lower frequency, but as the material is purified its resistivity increases and it is more difficult to couple sufficient power into the ingot to melt a zone. The reason for this can be understood by considering the variation of the skin depth (d) with resistivity (ρ) and frequency (f); d is proportional to $\sqrt{(\rho/f)}$ and is numerically equal to 8 mm for 0·01 Ω cm material and 400 kc/s. For optimum coupling the skin depth should be less than half the radius of the ingot (assumed cylindrical); in a typical case the mean radius of an ingot would be about 4 mm. Consequently for material of the above resistivity and higher, 5 Mc/s is much better than 400 kc/s. It is possible to overcome this kind of difficulty when using a 400 kc/s generator by preheating the ingot to 1000°C or more, in one of the furnaces. This raises the conductivity and consequently the coupled power.

Whenever radio-frequency heating is used it is found that a deposit of GaAs accumulates gradually on the container walls. It is believed that this transport from the liquid zone is due to the vigorous escape of arsenic from the centre of the zone where the temperature is well above the melting point of GaAs and the dissociation pressure exceeds the ambient. Small particles of gallium arsenide are ejected from the melt and impinge on the nearby walls.

The radiation losses from a zone heated by induction are quite large, so that the freezing front is concave towards the liquid. This applies especially to the arrangement of Fig. 3 and the effect is to make the material markedly polycrystalline. Radiation heating is generally an improvement in this respect at the expense of efficient stirring of the liquid.

(d) *Crystal growing*

The techniques of crystal growing which have been used successfully in the case of germanium and silicon may in principle be applied to gallium arsenide with suitable modification to provide the ambient arsenic pressure. Crystals of GaAs and InAs have been grown by the Czochralski method. GREMMELMAIER[11] has overcome the difficulty of controlling the movement of a seed crystal within a sealed system by the use of a magnetic lifting mechanism, while RICHARDS[12] has accomplished the same thing in a rather different way by using an annular liquid-gallium trap separating the moving and the stationary portions of the equipment. This latter method is ingenious but there are some difficulties in the operation.

Perhaps the most potentially attractive method for GaAs is the one described by CRESSELL and POWELL[13] and BENNETT and SAWYER[14] for the particular case of germanium. Here, the crystal is grown in a horizontal boat from an orientated seed by the controlled traversal of the liquid–solid interface along the ingot. The apparatus required is essentially simple and no difficult manipulations are involved. Although a certain amount of success has been achieved with this method, there are difficulties encountered with GaAs which do not occur for germanium. The first of these arises in the choice of crucible material. Quartz would be the obvious choice but unfortunately there is a tendency for localized "wetting" to occur. A partial solution is obtained quite simply by grinding the interior of the quartz boat with fine carborundum and coating with a deposit of pure carbon. For example, "AnalaR" acetone, when burnt, provides a source of suitably pure carbon. This is rubbed into the ground quartz surface and vacuum-baked. It is found, however, that GaAs is contaminated whenever the boat is treated in the above manner. There is some evidence that the cause of the contamination is not any impurity in the carbon but the carbon itself. Four such ingots had *n*-type Hall constants of 15, 44, 7·4 and 6·3 $cm^3\ C^{-1}$ and one doped with spectrographically pure carbon had a Hall constant of 8·1 $cm^3\ C^{-1}$. These are an order of magnitude worse than the values for the average GaAs prepared. HANNAY[15] has reported the presence of 3 p.p.m. of carbon in an ingot of gallium arsenide after extensive zone refining in a graphite boat. This is an order of magnitude less than the concentration corresponding to the Hall constants mentioned above. A second difficulty which occurs is common to the crystallization of all compounds possessing a highly volatile component. Fluctuations of the ambient arsenic pressure around the value 0·9 atm can cause bubbles of arsenic vapour to form within the liquid. These can be trapped by the movement of the advancing interface and incorporated in the solid.

The boat and contents, together with excess arsenic, are placed in a long quartz tube, evacuated and sealed (Fig. 4). The furnace system mounted

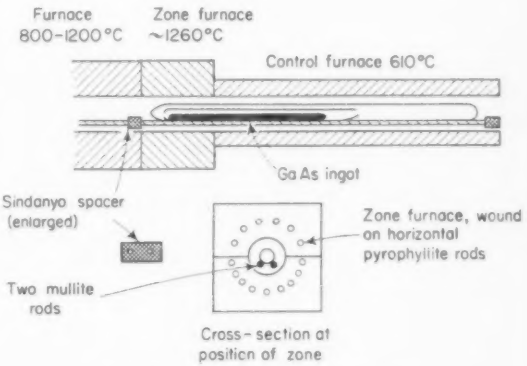


FIG. 4. Apparatus for crystal growing.

on a movable carriage consists of three separate parts. The first part is kept at 610°C, the minimum temperature of the system, and controls the pressure in the tube. At the beginning of the experiment one end of the ingot is situated in the second region at 1260°C and a molten zone is formed. The ingot is moved slowly into the third region in which the power input is set to produce a suitable temperature gradient at the freezing front. The centre furnace is wound in such a way that the power supplied to the upper and lower

halves can be regulated separately. This makes provision for adjusting the inclination of the freezing front to the vertical. It is possible to obtain very large crystals even in the absence of a seed, Fig. 2(b).

The following conditions are recommended for the successful application of this technique: (1) A constant arsenic vapour pressure of 0·9 atm should be maintained. (2) The freezing front should be planar and normal to the growth direction. A small temperature gradient assists in obtaining a planar interface.[13] (3) The temperature of any point in the liquid zone should not be much in excess of the melting point. This prevents significant unbalance between the dissociation pressure of the liquid and the ambient. (4) A slow rate of growth should be used.

The most satisfactory method used by the authors for growing crystals of GaAs has been the floating-zone technique commonly used in silicon technology. The application of this technique to gallium arsenide has been described by WHELAN and WHEATLEY[4]. Fig. 5 shows schematically the equipment which the present authors have used. A square-section rod of GaAs is cut from a cast ingot and fitted to a quartz chuck by careful grinding of its edges. A seed crystal is fitted to the lower chuck and held in position by a constriction in the quartz, while the GaAs rod is supported by a silica pin sealed through the upper chuck. A small separation of seed and rod is necessary to allow expansion of the material when heated to the temperature of the ambient furnaces. Excess arsenic is added to the system before evacuation. Silica rods are fused to either end of the assembly, the lower one being supported by a vertical pillar motor-driven at a rate of between 4 and 5 in/hr and the upper one passing through a guide to keep the assembly vertical. The furnaces are separated by a gap of approximately 0·75 in. in which a three-turn work coil of an 8 Mc/s r.f. generator is situated. Annular baffles, and a cylindrical mica window between the furnaces, are employed to reduce convection currents through the furnaces.

The zone is initially formed at the butt joint between the rod and the seed. Having produced a zone such that the tangent to the lower solid–liquid interface is vertical, it is traversed up the rod. Fig. 6 shows an example of a crystal grown in this way. The seed crystal was cut from a large crystallite in a horizontally cast ingot. The axis was found to be a (321) direction.

This method has the following advantages. Firstly there is no contact between the molten material and the silica container. Secondly, the agitation produced within the zone by the circulating r.f. currents leads to homogeneous material free from arsenic vapour "pockets".

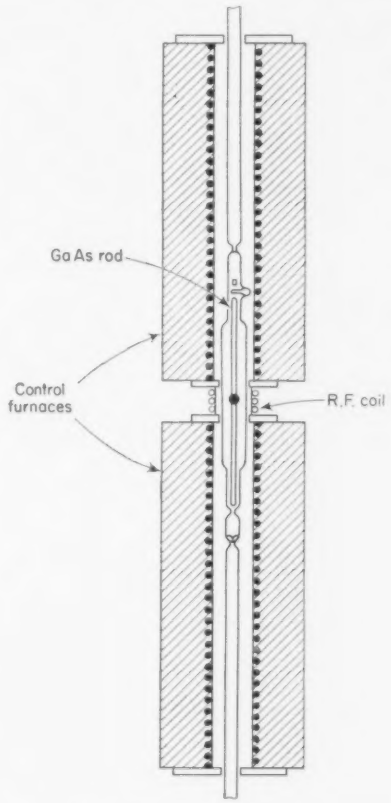


FIG. 5. Floating-zone equipment.

The principal disadvantage is the limitation in the diameter of the crystal which may be grown. The maximum length of a stable zone has been calculated by HEYWANG[16]. This depends on both the surface tension of the liquid and its density. Increasing the diameter of the rod while preserving a zone length less than the maximum value leads to the practical problem of obtaining a complete melt-through. The surface tension of liquid

GaAs has not been measured but the greater density of GaAs compared with silicon makes it rather less suited than silicon to this method.

In addition to the use of this technique for the growth of single crystals, zone purification by repeated passes has also been attempted. While this operation has obvious advantages over the more conventional zone purification described previously, the present apparatus is not ideally suited to it for the following reasons. Firstly, there is no facility for adjustment of the length of the zone and slight variations in the diameter of the zone tend to be accentuated by successive passes. It has been shown that this difficulty can be overcome in a more recent equipment by introducing a cylindrical sliding-joint fabricated in quartz. This consists of a quartz rod approximately 16 mm in diameter which is ground and polished over a length of 6–7 cm to be a close fit (approximately 10μ clearance on the diameter) in a quartz tube of similar internal surface finish. This improvement makes the equipment demountable and at the same time a negligible amount of arsenic is lost during operation. Secondly, there is a gradual deterioration in visibility through the silica envelope owing to material depositing on the inner wall. The greatest number of passes accomplished has been five.

(e) *Purity of gallium arsenide*

Although gallium and arsenic have been obtained from several sources and treated in a variety of ways prior to making the compound, the results may be considered, broadly, under two headings: (1) for GaAs, prepared and directionally frozen as described in Subsection (b). (2) for GaAs, which has been zone-refined.

(1) A number of ingots of GaAs have been made recently using gallium produced by Aluminium-Industrie-Aktien-Gesellschaft. This was used without further treatment while the arsenic from the American Smelting and Refining Co. was sublimed and treated with lead as described earlier. Specimens cut from these ingots at the first end to freeze have had a Hall constant (R) from -100 to $-1200\ \mathrm{cm^3\,C^{-1}}$ and an electron mobility ($R\sigma$) from 4000 to 6700 $\mathrm{cm^2\,V^{-1}\,sec^{-1}}$, the best value, at room temperature.

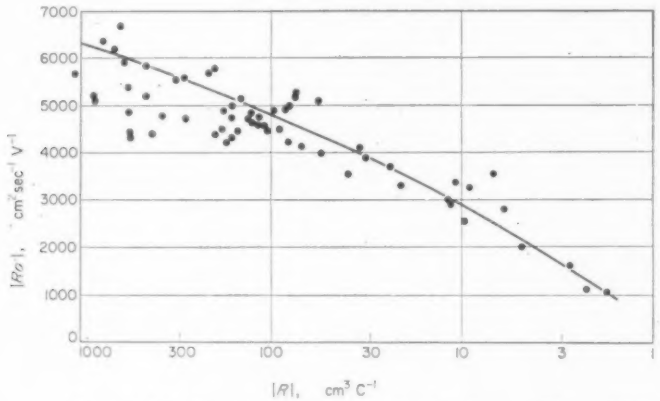


FIG. 7. Mobility of electrons in GaAs of various purities at 20°C.

A typical example had an electron mobility of 9000 $\mathrm{cm^2\,V^{-1}\,sec^{-1}}$ at 90°K. Specimens from the last end to freeze usually had higher resistivity, sometimes by several orders of magnitude. It is believed, however, that in these cases the number of carriers was not an indication of the total impurity content (probably $\sim 10^{17}/\mathrm{cm^3}$). It is likely that the high resistance arises either through the compensation or over-compensation of donor by acceptor impurities with relatively large ionization energy, possibly oxygen or copper.

(2) In the case of horizontally refined ingots the values of R were all negative and smaller than those described above in (1). This was taken to mean that the GaAs contained donor and acceptor

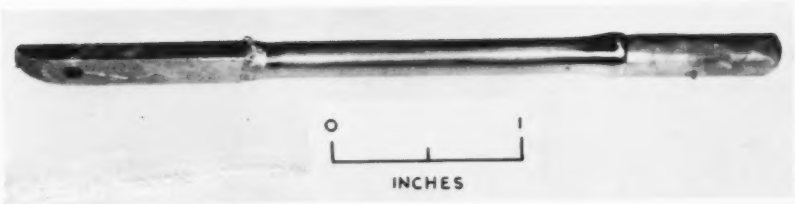


FIG. 6. Gallium arsenide single crystal.

[*facing p.* 102

impurities and that the latter were swept more readily to the last end to freeze. As a result the total impurity content would be lower over the majority of the ingot than if treated by method (1). One serious inconsistency, which has not yet been resolved, is that material made by process (1) shows higher values of electron mobility than that made by process (2). This might happen if impurities in material (1), donor and acceptor, were associated to make uncharged pairs with a small scattering cross-section.

The results obtained on ingots which were refined by the floating-zone method show some similarity to those discussed above [both (1) and (2)]. The apparent segregation of impurities is better in this experimental arrangement than when the ingot is zone-refined horizontally, but it is believed that this is due to less contamination from boat and walls. One significant advantage of any method employing r.f. heating is the improved homogeneity which results from the agitation of the melt.

A curve is shown in Fig. 7 correlating Hall constant and electron mobility for a variety of specimens measured. While it is believed that the intrinsic Hall constant in the range 10^2–10^3cm^3 C^{-1} cannot be used as a measure of the total impurity content in the specimens, it is significant that $R\sigma$ increases with R. If the total impurity content is in the region of 10^{17} cm^{-3} or greater it seems unlikely that the compensation of donors and acceptors could be a random process. It is possible that some form of automatic self compensation takes place.

During the work, a considerable amount of spectrographic information on impurities in gallium and arsenic has been gathered. Copper, silver, magnesium, calcium and lead have been most frequently detected in gallium; and copper, silver, magnesium, calcium, sodium, iron and silicon in arsenic. Such elements as oxygen, sulphur and selenium, which are not detected by spectrographic techniques, are believed to be present in one or the other.

3. MEASUREMENTS AND TECHNIQUES

(a) *Electrical properties of impurities*

As is well known, the electrical behaviour of semiconductors depends on the amount and nature of the impurity content. Observations on this behaviour are often useful in the identification of impurities and indicate which elements may be suitable for the purpose of doping. Consequently, a study was made using the known impurities in GaAs, i.e. those detected by spectrograph in gallium and arsenic, and certain other elements such as Zn, Sn, Si and Te. Silicon was of special importance since the possibility of pick-up of this element from quartz-ware must always be borne in mind.

Work on similar lines involving impurities in InSb, InAs, GaSb and GaAs has been described by EDMOND[17]. The method employed to prepare the material was to melt about 5 g of relatively pure GaAs in a small evacuated quartz tube with a few milligrammes of impurity and a little excess arsenic.

The material was kept in a liquid state at 40°C above the melting point for 1 hr and then cooled slowly. Single-crystal sections were usually formed of sufficient size to provide rectangular specimens for measurements of R and σ. The results are shown in Table 1.

The seventh column is calculated from the weight of impurity and of the compound on the basis of one carrier per impurity atom and assuming that $R = 1/ne$ where n is the number of carriers per cm^3 and e is the value of the electronic charge. The measured and calculated values of R differ not only because of solubility considerations but also as a result of some segregation during solidification.

The results show some interesting features: (1) Copper is an acceptor impurity, as found also by FULLER and WHELAN[18]. Silver and gold seem to be much less soluble than copper. (2) Magnesium behaves as a donor impurity. This rather unexpected behaviour could be due to magnesium atoms occupying an interstitial position in the lattice. (3) Silicon is a donor impurity. Recently KOLM *et al.*[19] have shown that silicon is soluble in GaAs up to 0·5 per cent and they suggest that silicon atoms substitute for nearest-neighbour pairs. The fact that the present work shows that silicon is a donor impurity implies that silicon atoms may also replace single gallium atoms. (4) Iron is an acceptor impurity. Measurement of resistivity over a range of temperature shows an ionization energy of 0·37 eV.

Table 1

Before Doping		After Doping					
R (cm^3/C^{-1}) (290°K)	$\rho \times 10^4$(Ωcm) (290°K)	Impurity added	R (cm^3/C^{-1}) (290°K)	$\rho \times 10^4$(Ωcm) (290°K)	R (cm^3/C^{-1}) (90°K)	Calculated R (cm^3/C^{-1})	Impurity type
−38	300	Cu	+883	68,300		0·39	*p*
−48	280	Cu	+94	3400	+581	0·55	*p*
−120		Au	−78			1·03	?
−86	280	Ag	−89	520	−96	0·52	?
−86	280	Mg	−5			0·12	*n*
−86	280	Mg	−2·3			0·11	*n*
−152	479	Mg	−2·7	16·5	−2·3	0·1	*n*
−86	280	Ca	−17·1	61	−17·1	0·23	(*n*?)
−75		Zn	+1·1	110		0·39	*p*
−86	280	Si	−2·5		−2·7	0·13	*n*
−86	320	Si	−1·5	21		0·16	*n*
−35		Sn	−13	60	−21	0·62	(*n*?)
		S	−1·4			0·24	*n*
−80		Te	−2·8			0·86	*n*
−72	199	Te	−1·8			0·045	*n*
−53		Pt	−10	48	−63	0·80	
−88		Fe	−301	1800	−537	0·29	(*p*?)
−310	967	Fe	+9800			0·11	*p*

(b) *Heat treatment of GaAs*

In the course of measurements of impurity diffusion in GaAs it has been found that the electrical properties of the material, normally *n*-type, may suffer a pronounced change as a result of heat treatment. The treatment has consisted of heating mostly at 1000°C for 1–60 hr in a small evacuated capsule of quartz (Vitreosil) and cooling rapidly to room temperature. The modification to the electrical properties has always been such that the electron concentration has been reduced and in some cases the type of conductivity has changed from *n*-type to *p*-type. A number of experiments have been carried out to investigate these changes and the conclusion reached is that in a number of cases they are due to contamination by copper. The copper has been identified in the GaAs by means of a square-wave polarograph and by comparing resistivities as a function of temperature of heat-treated specimens and copper-doped specimens, all *p*-type. Furthermore it is found that the changes are an order of magnitude smaller if the purest quartz (Spectrosil) is used instead of Vitreosil, e.g. in Spectrosil the change in carrier concentration in the extrinsic range Δp is estimated to be about 10^{16} cm^{-3} for 15 hr at 1000°C and in Vitreosil $1–5 \times 10^{17}$ cm^{-3} for the same time and temperature. The effect also depends on the duration of the heating, e.g. Δp is equal to 10^{17} cm^{-3} for 1 hr and $7·5 \times 10^{17}$ cm^{-3} for 60 hr at 1000°C (Vitreosil) for specimens from the same ingot. A tendency for the specimens to return towards their original state on annealing has been noted. This effect would be consistent with the precipitation of copper at dislocations as suggested by Dixon and Enright[20] in the case of InAs.

It is hoped to publish these results later in more detail.

(c) *Preparation of contacts*

Low-resistance ohmic contacts to both *n*-type and *p*-type GaAs are required for routine electrical measurements and in the fabrication of devices. The method to be employed is determined partly by the area of contact required. The following techniques have been used successfully.

(1) Pure tin or indium heated to between 400 and 500°C in contact with clean *n*-type material in an inert atmosphere produces a strong mechanical join. This is suitable for end contacts to Hall

specimens or base contacts to rectifiers. Indium, doped with a small percentage of zinc (4·8 atomic per cent zinc), is suitable for *p*-type material.

(2) Silver electroplating can be used for end contacts to both *p*-type and *n*-type material. For *p*-type samples one layer of silver is usually sufficient but for *n*-type specimens the technique employed is to plate, discharge a condenser (say $0{\cdot}25\mu$ F at a suitable voltage) between the ends of the specimen, and plate again.

(3) Gold wires "formed" either by a condenser discharge or by the passage of an a.c. current are useful for small-area Hall or resistivity probes on both *n*-type and *p*-type material. Bulk contamination of the specimen with gold is unlikely since it is believed that the diffusion of gold is slow in GaAs ($D \sim 2{\cdot}5 \times 10^{-11}$ cm^2 sec^{-1} at 1000°C[21]).

(4) Fine gold wires can be bonded to a clean GaAs surface by the combined application of pressure and heat. The temperature required is well below the melting point of gold. This type of contact is especially useful where the contact has to be very restricted in area. Alternatively, fine gold wires (0·002–0·003 in. dia.) can be alloyed with a clean GaAs surface heated to a temperature of 550–600°C in an inert atmosphere.

The range of application of each of these methods could be extended by suitable modifications of the experimental technique.

(d) *Etching*

SCHELL[22] has discussed preferential and non-preferential etches for GaAs, and their application to the display of dislocations.

When it has been necessary to clean the surface of GaAs the authors have found that a warm mixture of 3 volumes concentrated sulphuric acid, 1 volume of water and 1 volume of hydrogen peroxide leaves a polished finish;* a layer of GaAs about 0·0005 in. thick is removed in 2 min. A good alternative is a hot mixture of concentrated hydrochloric acid and nitric acid (approximately five to one by volume).

A less reactive mixture than the above, consisting of 1 volume sulphuric acid, 8 volumes of water and 1 volume of hydrogen peroxide, is valuable for showing grain structure.

Hydrochloric acid and nitric acid, cold, in the ratio of five to one, with or without a small amount of water, can be used to detect the position of a *p–n* junction.

(e) *Diffusion*

p–n junctions have been prepared by diffusion of zinc from the vapour phase into *n*-type GaAs. Attempts to measure the diffusion constant of zinc in GaAs by comparison of junction depths in material of two different electron concentrations have been unsuccessful. Subsequently radio-tracer techniques have established that the zinc diffusion process does not follow the simple concentration-independent laws at the concentrations examined and have thereby explained the failure to obtain satisfactory results from the electrical method.[23]

Acknowledgements—The authors wish to thank Mr. N. A. C. THOMPSON and Mrs. I. S. PINNOCK for making some of the measurements, Miss J. GOODSON for preparing some of the material, Dr. O. SIMPSON for assistance in preparing the manuscript, and the Admiralty for permission to publish the work.

REFERENCES

1. O. G. FOLBERTH and H. WEISS, *Z. Naturf.* **10**a, 615 (1955).
2. J. T. EDMOND, R. F. BROOM and F. A. CUNNELL, *Report of the Meeting on Semiconductors, Physical Society, London, April* 1956, p. 109.
3. J. VAN DEN BOOMGAARD and K. SCHOL, *Philips Res. Rep.* **12**, 127 (1957).
4. J. M. WHELAN and J. H. WHEATLEY, *J. Phys. Chem. Solids* **6**, 169 (1958).
5. J. R. RICHARDS, *Nature, Lond.* **177**, 182 (1956).
6. D. P. DETWILER and W. M. FOX, *J. Metals, N.Y.* **7**, 205 (1955).
7. E. SCHILLMAN, *Z. Naturf.* **11**a, 463 (1956).
8. T. C. HARMAN, E. P. STAMBAUGH and H. L. GOERING, *Semiconductor Symposium of the Electrochemical Society, Cleveland, October*, 1956.
9. J. M. WHELAN, J. D. STRUTHERS and J. A. DITZENBERGER. Private communication.
10. J. VAN DEN BOOMGAARD, F. A. KROGER and H. J. VINK, *J. Electronics* **1**, 212 (1955).
11. R. GREMMELMAIER, *Z. Naturf.* **11**a, 511 (1956).
12. J. L. RICHARDS, *J. Sci. Instrum.* **34**, 289 (1957).
13. I. G. CRESSELL and J. H. POWELL, *Progress in Semiconductors*, Vol. II, p. 37. Heywood, London (1957).
14. D. C. BENNETT and B. SAWYER, *Bell Syst. Tech. J.* **35**, 637 (1956).

* This etch was suggested by N. RAIN of this laboratory.

15. N. B. HANNAY, *Semiconductors* p.413. Reinhold, New York (1959).
16. W. HEYWANG, *Z. Naturf.* **11**a, 238 (1956).
17. J. T. EDMOND, *Proc. Phys. Soc. Lond.* **73**, 622 (1959).
18. C. S. FULLER and J. M. WHELAN, *J. Phys. Chem. Solids* **6**, 173 (1958).
19. C. KOLM, S. A. KULIN and B. A. AVERBACH, *Phys. Rev.* **108**, 965 (1957).
20. J. R. DIXON and D. P. ENRIGHT, *J. Appl. Phys.* **30**, 573 (1959).
21. H. B. CLARKE. Unpublished work at Services Electronics Research Laboratory (1959).
22. H. A. SCHELL, *Z. Metalk.* **48**, 158 (1957).
23. J. W. ALLEN and F. A. CUNNELL, *Nature, Lond.* **182**, 1158 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 107–122. Printed in Great Britain

THE BEHAVIOR OF *p*- AND *n*-DOPED CONTACTS IN A SPACE-CHARGE DEPLETION REGION*

J. M. LAVINE

Research Division, Raytheon Company, Waltham 54, Massachusetts

Abstract—A qualitative description of the floating potential, the forward and the reverse characteristics of both *p*- and *n*-doped contacts located in the space-charge depletion region of a reverse-biased *p*–*n* junction is presented. A model to explain the occurrence of space-charge-limited injection of electrons from an *n*-doped contact and of holes from a *p*-doped contact is described. A method of obtaining increased injection of electrons from a *p*-doped contact is shown. By analogy with a vacuum diode this mode of injection is termed "temperature-limited". Measurements of the floating potential of both *p*- and *n*-doped contacts and of the space-charge-limited and temperature-limited currents are reported.

Résumé—On donne une discussion qualitative sur les caractéristiques directes et inverses des contacts dopés *n* et *p* situés dans la région de charge d'espace d'une jonction de polarisation inverse. On donne un modèle pour expliquer l'injection d'électrons ou de trous limités par la charge d'espace. On peut obtenir une augmentation de l'injection des électrons d'un contact dopé *p*. Par analogie avec les diodes à vide ce type d'injections est appelé "temperature-limited".

Zusammenfassung—Es wird eine qualitative Beschreibung des Floating-Potentials, der Fluss- und der Sperr-Charakteristik sowohl von *p*- als auch *n*-dotierten Kontakten gegeben, die im Verarmungsgebiet der Raumladungszone eines in Sperr-Richtung geschalteten *p*–*n*-Überganges angebracht sind. Ein Modell wird beschrieben, welches das Auftreten einer raumladungsbegrenzten Injektion von Elektronen aus einem *n*-dotierten Kontakt und von Löchern aus einem *p*-dotierten Kontakt erklärt. Eine Methode wird aufgezeigt, mit deren Hilfe man eine verstärkte Injektion von Elektronen aus einem *p*-dotierten Kontakt erhält. In Analogie zur Vakuum-Diode wird diese Art der Injektion eine "temperaturbegrenzte" genannt. Über Messungen des Floating-Potentials sowohl der *p*- als auch *n*-dotierten Kontakte und der raumladungsbegrenzten sowie temperaturbegrenzten Ströme wird berichtet.

1. INTRODUCTION

In the near vicinity of a reverse-biased *p*-*n* junction there exists a space-charge depletion region with electric fields of the order of several thousand V/cm. The width of the depletion region depends upon the impurity concentration in the vicinity of the junction, the dielectric constant of the material, the applied voltage and in certain cases upon the geometry of the structure. The space-charge potential distribution and the width of the space-charge region can be readily calculated[1] for a plane parallel geometry (see Fig. 1) provided that the impurity distribution on both sides of the junction is known.

We have been concerned mostly with silicon structures of the form shown in Fig. 1. In the plane parallel structure, space-charge depletion regions ranging from 2 mils to 20 mils have been obtained with about 400 V of reverse bias, using material varying from 10 Ω-cm to 1000 Ω-cm. The *p*-*n* junctions have been fabricated by deep diffusions which approximate a linear graded structure near the junction.

The lower portion of Fig. 1 shows a beveled structure used to obtain an increase in the width of the space-charge region near the surface. Beveling provides a geometrical increase of space-charge region at the surface as well as an increase due to the space-charge neutrality requirement on both sides of the junction. The number of positive charges on the *n* side of the junction must equal

* Supported in part by BuShips under Contract No. NObsr 72787.

the number of negative charges on the p side of the junction. Since the density of charge on the low-resistivity side of the junction is small, and

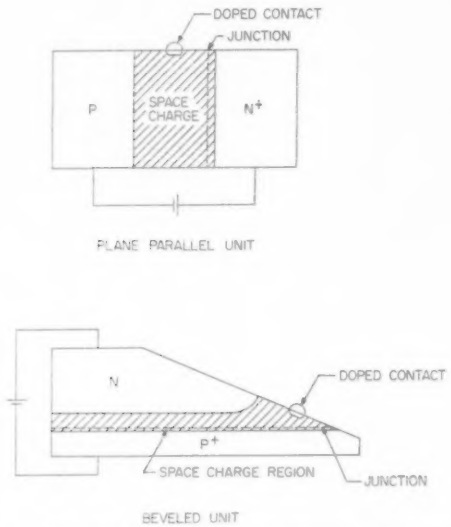


FIG. 1. Two geometries used for the measurements.

because the electric field must be normal to the surface in the absence of surface charges, there is a considerable increase of space-charge region near the surface of the bevel.

Small n- and p-doped contacts of the order of 1 to 2 mils in diameter have been alloyed into the space-charge region of both the plane parallel and beveled structures using standard transistor techniques. Similar contacts have been diffused into these regions. This paper is concerned with describing the current–voltage relationships of such contacts, with describing their floating potential and with suggesting a method for improving the forward injection properties of such contacts.

2. THE MEASUREMENT TECHNIQUE

When several hundred volts of reverse bias are applied to the diode structures of Fig. 1, the small n- or p-doped contacts lie wholly within the space-charge region. These contacts will float at a potential which depends upon their location in the space-charge region, their size and the main-diode voltage. Electrical connection to these small doped contacts was made with a micromanipulated tungsten point. The floating potential of the contacts was measured with a Kay Lab Model No. 202B Voltmeter with an input impedance of 100 MΩ.

Potential profiles of the space-charge region were obtained by micromanipulating a tungsten point, measuring distance by means of a calibrated microscope eyepiece and measuring voltage with the Kay Lab Voltmeter.

The sign of the charge carrier injected into a space-charge region can be unambiguously determined by using the circuit shown in Fig. 2. In

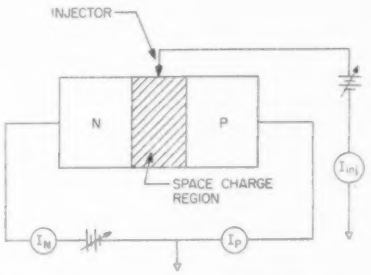


FIG. 2. Circuit for measuring injected-electron or -hole current.

Fig. 2 is shown a metal point-contact injector which is capable of injecting either electrons or holes into the space-charge region. It was first shown by MATTHEI and BRAND[2] and later by STATZ *et al.*[3] that a metal point injects holes if biased positively with respect to its floating potential and injects electrons if biased negatively with respect to its floating potential. Injected electrons enter the space-charge region, flow toward the n side of the main diode, through the ammeter marked I_N, to ground and back through the injecting circuit. Injected holes, on the other hand, flow to the p side of the main diode, through the ammeter marked I_P, to ground and then back through the injecting circuit. The ammeter in the injecting circuit will change sign with change of sign of the injected carrier. The ammeters I_N and I_P will read the injected-electron or -hole currents respectively (and the very small leakage current of the main diode). The combination of the three meters permits an unambiguous determination of the sign of the injected carrier even when the possibility of charge collection at the small doped contact occurs.

3. RESULTS

(a) *Typical current–voltage characteristics*

Fig. 3 shows typical characteristics obtained from a metal injecting contact. The origin of the current–voltage characteristic is the floating potential of the point. The notable characteristics of metal contacts in the space-charge region are (1) the relatively modest currents obtained with relatively large voltages (applied above the floating potential) and (2) the symmetry of injecting either electrons or holes. The maximum forward conductance observed with point-contact injectors has been of the order of 100 μA/V. Extrapolating from this behavior to a doped contact we would expect to observe an improved forward conductance and a reverse characteristic similar to that of a high quality p–n junction. In the forward direction, an improvement over the point contact would be expected because of the larger area of the contact and the more intimate contact of an alloyed structure. In the reverse direction, a good quality characteristic is expected because the current that flows will consist (in the case of a p-doped contact) of a small portion of the holes collected from the main-diode saturation current, the holes and electrons generated in the vicinity of the contact, and the electron diffusion component of the saturation current of the small contact. Since the first component will be less than the contribution of normally neutral material, the sum of the three components is expected to be less than that of a small-area reverse-biased diode. Experimentally, no marked improvement in the forward conductance over that observed for point contacts was observed for either p-doped or n-doped contacts, and only infrequently were good reverse characteristics obtained. Fig. 4 shows the current–voltage characteristics of a p-doped and an n-doped contact in a space-charge region. The curves shown in Fig. 4 are data of contacts which displayed good reverse characteristics. We note that the forward conductance of these doped contacts is not markedly improved over that of the point contacts.

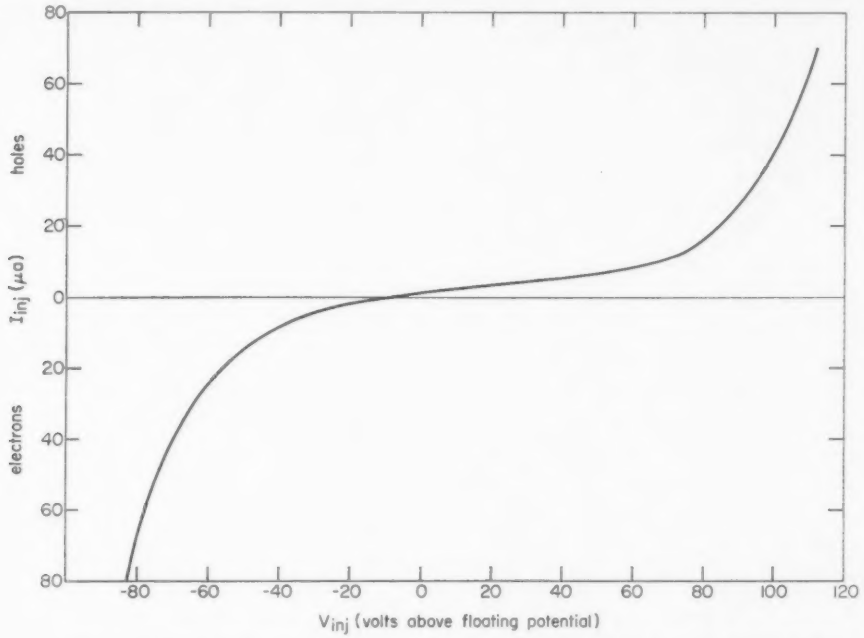
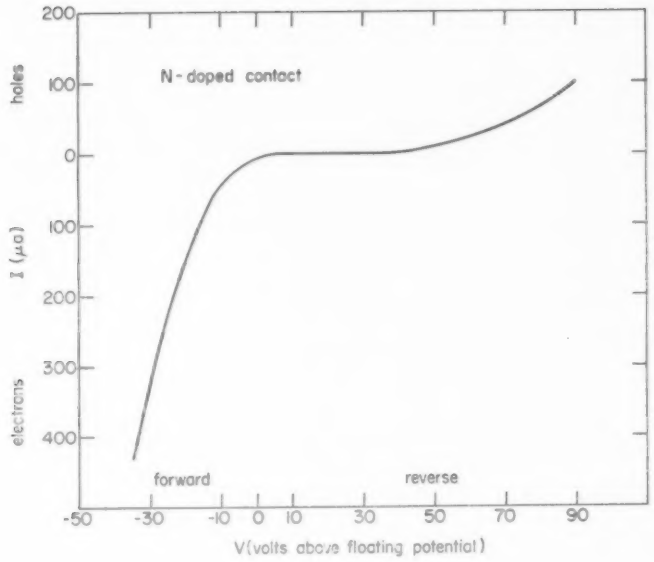
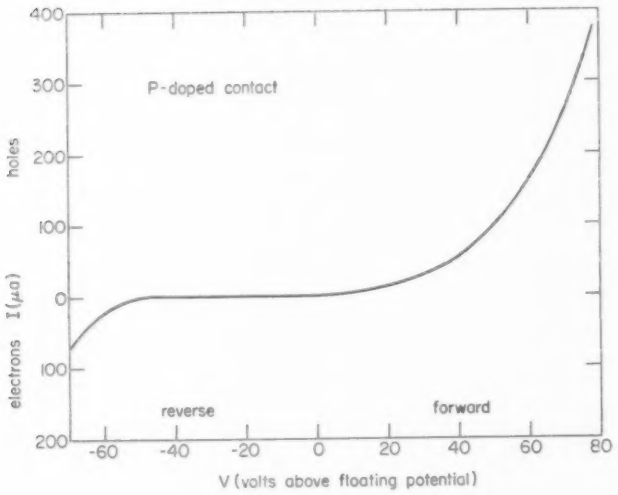


FIG. 3. Typical characteristics of a point contact injecting into a space-charge region.

In order to test whether our small n- and p-doped contacts were good quality contacts, measurements were made with the space-charge region collapsed to display an underlying neutral material of the opposite conductivity type from



(a)



(b)

FIG. 4. Typical characteristics of *n*-doped and *p*-doped contacts in a space-charge region.

the small contact. Fig. 5 shows the forward characteristic of a small p-doped contact inside (lower figure) and outside (upper figure) the space-charge region. Outside the space-charge, injected

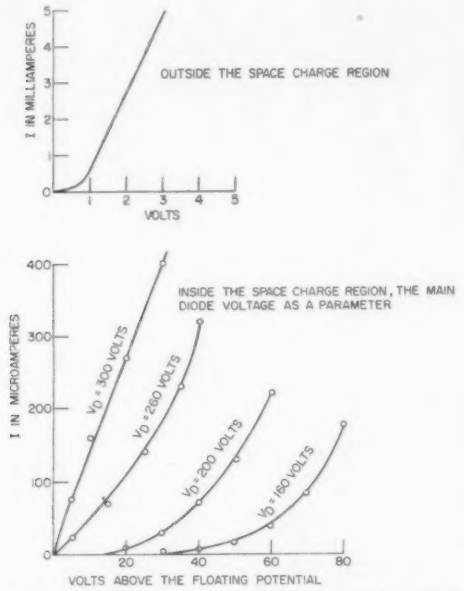

I IN MILLIAMPERES
OUTSIDE THE SPACE CHARGE REGION
VOLTS
I IN MICROAMPERES
INSIDE THE SPACE CHARGE REGION, THE MAIN DIODE VOLTAGE AS A PARAMETER
V_D = 300 VOLTS
V_D = 260 VOLTS
V_D = 200 VOLTS
V_D = 160 VOLTS
VOLTS ABOVE THE FLOATING POTENTIAL

FIG. 5. Forward characteristics of a p-doped contact inside and outside of the space-charge region.

current of the order of 4 mA was obtained for voltage of the order of 3 V (forward conductance of the order of 1500 μA/V). The resistance in series with the diode is to be expected from the diameter of the contact and the resistivity of the high-resistivity material. The same contact inside the space-charge region shows an injected current (of holes) of about 100 μA for about 80 V applied above the contact floating potential (forward conductance of the order of 10 μA/V). It is observed, however, that with increasing voltage across the main diode (and between the p-contact and the p region of the main diode), the forward conductance of the contact is improved by a substantial factor, but the forward conductance is still only about 15 μ A/V. This behavior, observed for all contacts injecting into a space-charge region, is probably due to the space-charge-limited character of the current flow. A model to explain this behavior will be discussed below.

(b) *The floating potential*

Fig. 6 shows the assumed potential distribution about a p-doped contact floating (no applied bias) in the space-charge region. In order to simplify the description, the equipotential lines far from the contact have been drawn equally spaced, although

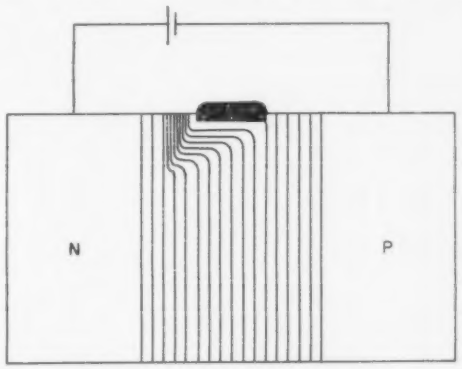


FIG. 6. Assumed potential distribution about a p-doped contact in the space-charge region.

such a potential distribution cannot be achieved around a p–n junction. The p-doped contact is a relatively high conductivity contact and, therefore, assumed to be an equipotential surface. Because of the difficulty in solving Poisson's equation subject to this boundary condition, no solution has been obtained for a perturbation such as this upon the space-charge potential. The effects of the equipotential p-doped contact on the potential at the edges of the space-charge region are also not known. For the present, we are mainly concerned with the region near the contact, and we ignore the effects at the boundaries although they are important in any quantitative treatment of the problem.

With no bias applied to the p-doped contact, either as many holes are collected by the contact as are injected, or no carriers pass between the two regions. We assume that the latter condition obtains. That is, a barrier tending to keep holes in the p-doped contact exists between the contact and the space-charge region. Alternatively stated, the p-doped contact tends to float so that it is everywhere reverse biased with respect to the underlying space-charge potential. This suggests that a p-doped contact will float at the lowest potential it sees in the space-charge region, at a potential such that no hole current will be injected. Fig. 7 shows

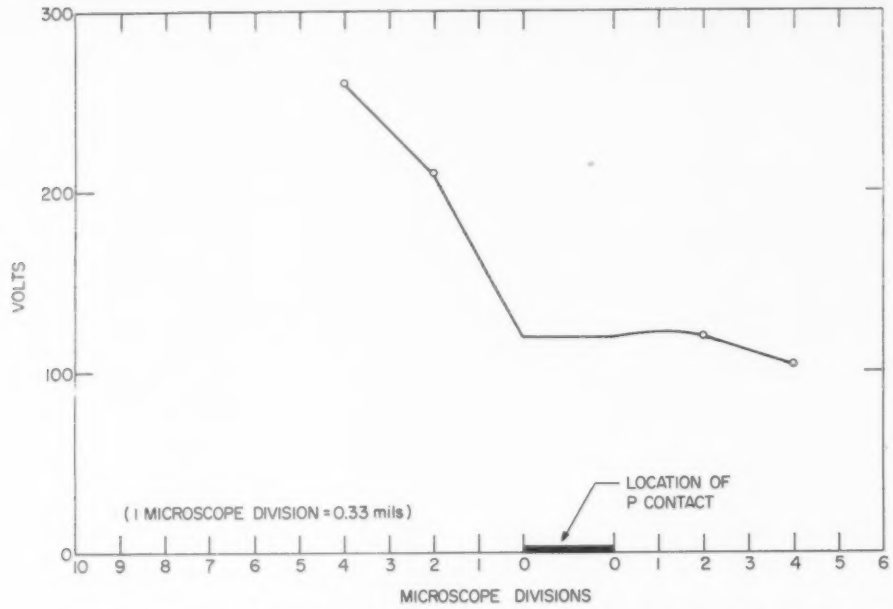


FIG. 7. Potential profile through a p-doped contact on a beveled chip. Main-diode bias is 400 V.

the measured potential through a p-doped contact floating in a space-charge region. In the absence of the contact, the space-charge potential is smooth through the region of the contact. The data indicate that the contact tends to float at the lowest voltage it encounters in the space-charge region. We also note that the slope of the potential at the high-voltage edge of the contact is greater than that which would exist in the absence of the contact. This increase of electric field at the high-voltage edge of the p-doped contact has been indicated in Fig. 6 by a crowding of the equipotential lines. We shall return to this increase of electric field in discussing the reverse behavior of such contacts below.

By analogy, we would expect that an n-doped contact would float at the highest potential it encounters in the space-charge region, and that the electric field at the low-voltage edge of the contact would be greater than that existing in the absence of the contact. Fig. 8 shows the potential measured through an n-doped contact located in a space-charge region. It is pointed out that the measurement of the potential as a function of distance is subject to some error because of the small spacings involved. The important features of the argument are, however, borne out by the measurements.

(c) *The reverse characteristic*

It is believed that the failure to observe good reverse characteristics in many units is attributable to the fact that the doped contacts distort the electric field in the vicinity of the contact (shown in Figs. 6 and 7 for a p-doped contact) in such a manner that avalanching fields may sometimes be present even with no voltage applied to the contacts. This effect can cause a large increase in the main-diode leakage current also, even with the doped contact floating. Removal of the doped contacts by etching has reduced the saturation current of the main diode in many cases. However, because of the uncertainty of the surface contribution to the main-diode leakage current this test is inconclusive. In order to reduce the electric field in the vicinity of the contact, it is necessary that the field in the absence of the contact be markedly reduced. Reduction of the space-charge electric field by beveling the chip or by using very high-resistivity silicon has resulted in small doped contacts with a very high yield of good reverse characteristics.

There appears to be some correlation between the diode breakdown outside and inside the space-charge region, in the plane parallel unit. Since the breakdown in both cases depends upon the electric field close to the p-doped contact, this is expected. In those cases where good reverse characteristics have been observed, negligible reverse current flows until breakdown is reached. When the p-contact breaks down, an electron current is observed in the space-charge region and a hole flow is observed out of the p-doped contact. Experimentally, this results in a reversal of the current flow in the injector ammeter from that observed in the forward direction, and an electron flow out of the n region of the main diode.

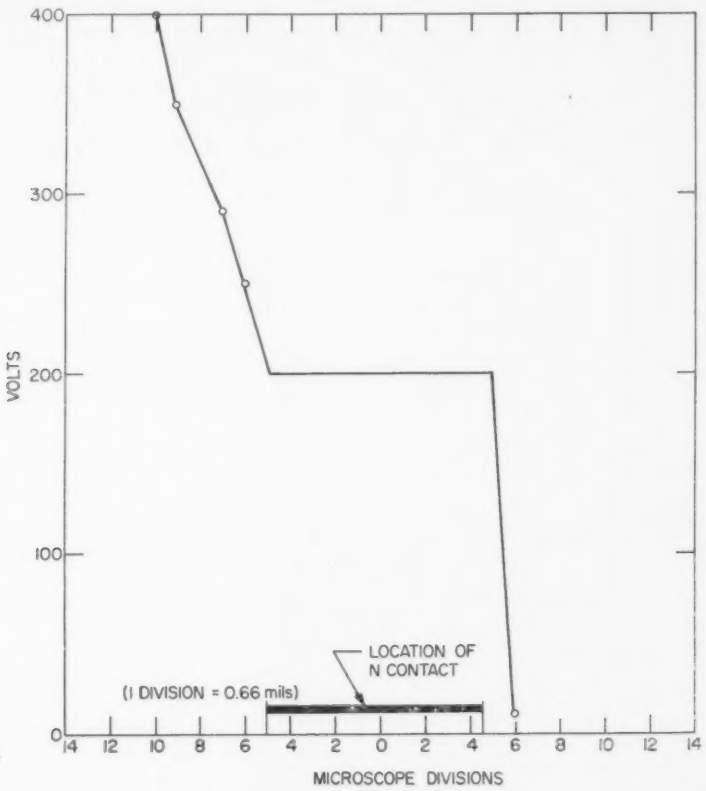


FIG. 8. Potential profile through an n-doped contact on a beveled chip. Main-diode bias is 400 V.

(d) *The forward characteristic*

(1) *Space-charge-limited injection.* In Fig. 5 the forward characteristics of a p-doped contact in the space-charge region have been shown with the main-diode voltage, V_D, as a parameter. Although the forward conductance of the contact increases with increasing V_D, it does not approach that of the contact on neutral n-type material. In order to account for this behavior it has been suggested that space-charge-limited current flows in the space-charge region.

It is the purpose of this section to present a plausibility argument for the existence of a space-charge-limited current flow from such contacts. As shown by SHOCKLEY and PRIM[4] the conditions prevailing in the space-charge region may be derived from Poisson's equation

$$\nabla^2\psi = -\frac{4\pi}{\epsilon}\left(\rho+\rho_f\right) \qquad (1)$$

where $\psi =$ the potential, $\rho =$ the mobile charge density, $\rho_f =$ the fixed charge density, and $\epsilon =$ the dielectric constant.

The current density J is given by

$$J = -D\nabla\rho + \mu\rho E \tag{2}$$

where D is the diffusion coefficient of the mobile charges, ρ, and μ is the mobility, related to D by the Einstein relation $D = \mu\, kT/q$. E is the electric field and $E = -\nabla\psi$. If we ignore the diffusion term, (2) reduces to

$$J = \mu\rho E = \rho v \tag{3}$$

where v is the carrier velocity. SHOCKLEY and PRIM have solved equations (1) and (3) for a one-dimensional n–i–n structure, subject to the conditions (i) diffusion effects can be ignored [i.e. equation (2) reduces to (3)], (ii) the effect of the fixed charges in the space-charge region can be ignored (i.e. $\rho_f \ll \rho$), (iii) the electric field is zero at $x = 0$, and (iv) the mobile carrier velocity is given by $v = \mu$E where μ is the carrier mobility. The solution they obtained is

$$J = \frac{9\mu\epsilon V^2}{32\pi w^3}. \tag{4}$$

where V is the applied voltage across the n–i–n structure and w is the width of the i region. This relationship was experimentally tested by DACEY[5] on a p–ν–p structure which allowed a one-dimensional analysis. DACEY pointed out that at very low current densities it was necessary to take into account the fixed charges in the space-charge region and at all current densities it was necessary to take into account the field dependence of the mobility. At high current densities subject to conditions (i), (ii) and (iii) above, plus the condition that the carrier velocity is given by $v = \mu(E_0 E)^{\frac{1}{2}}$ where μ is the low field mobility, E the electric field and E_0 the critical field at which the mobility becomes field-dependent, DACEY obtained the expression

$$J = \left(\frac{5}{3}\right)^{3/2} \frac{\epsilon\mu E_0^{1/2} V^{3/2}}{6\pi w^{5/2}}. \tag{5}$$

Finally, at very large fields the carrier velocity becomes saturated as shown by RYDER[6]. For the case in which a saturated carrier velocity v_s obtains, it can be shown, subject to the conditions (i), (ii) and (iii) above, that

$$J = \frac{\epsilon v_s}{2\pi w^2} V. \tag{6}$$

We thus observe that subject to the assumptions given above, the current density varies from a square, to a three-halves power, to a linear dependence upon the applied voltage provided that one mobility law holds for the region of current flow. An exact calculation would, of course, require that the complete dependence of mobility upon electric field be integrated over the path of current flow. This calculation would remove the contradiction that the electric field is equal to zero at the boundary while assuming in the latter two cases that the field is sufficiently large over the current path to use the high field-velocity.

In the case of the punch-through transistor structure examined by DACEY, the experimental data can be fitted quite well. However, in the case of the geometries used in these experiments several complications occur. First, it has not yet been possible to calculate the electric field between the doped contacts and either edge of the space-charge region. Even in the case of the plane parallel geometry, a one-dimensional approximation is invalid. Secondly, the application of an applied voltage distorts the potential in the space-charge region in a manner not readily calculated. This can be observed, for example, by measuring the potential distribution in the vicinity of a reverse-biased contact. This further complicates an already inexact knowledge of the electric field. Finally, when the p-doped contact is forward biased, injection first takes place at the low-voltage edge of the contact. Raising the potential of the contact (see Fig. 6) results in an increase in the area of the contact which is forward biased as well as an increase in the voltage of that region already injecting with respect to the space charge beneath. Therefore, the area of current-injecting contact is not constant, but depends upon the applied voltage. For these reasons it has not been possible to fit the experimental data with the expressions (4), (5) or (6). However, it will be made plausible below, that a boundary condition $E = 0$ does exist between a p-doped contact and the p region of the main diode giving rise to a space-charge-limited current. The details of this current depend upon a

better knowledge of both the electric field and the potential between the boundary $x = 0$ and the p region of the main diode.

We shall now show that it is reasonable to expect a boundary condition $E = 0$ between the p-doped contact and the p region of the main diode. Recent work by SAH *et al.*[7] suggests that under large reverse bias such as is present in the space-charge region, the situation is described by a quasi-Fermi level for electrons ϕ_n located in the valence band and a quasi-Fermi level for holes ϕ_p located in the conduction band. The relationship between the hole density and the quasi-Fermi level for holes is given by

$$p = n_i \exp q(\psi - \phi_p)/kT. \tag{7}$$

$\phi_p - \psi = \delta\phi_p$ represents the energy separation between ϕ_p and the center of the gap. From (7) we obtain

$$\delta\phi_p = \frac{kT}{q} \ln \frac{n_i}{p}. \tag{8}$$

Since the concentration of holes in the space-charge region is much less than n_i, $\delta\phi_p$ is a positive number. SAH *et al.* have estimated $\delta\phi_p$ by using the following expression for the current density:

$$J_p = -q\mu_p p \frac{d\phi_p}{dx}. \tag{9}$$

However, since it is believed that a saturated carrier velocity exists over most of the space-charge region we estimate $\delta\phi_p$ from

$$J_p = pqv_s \tag{10}$$

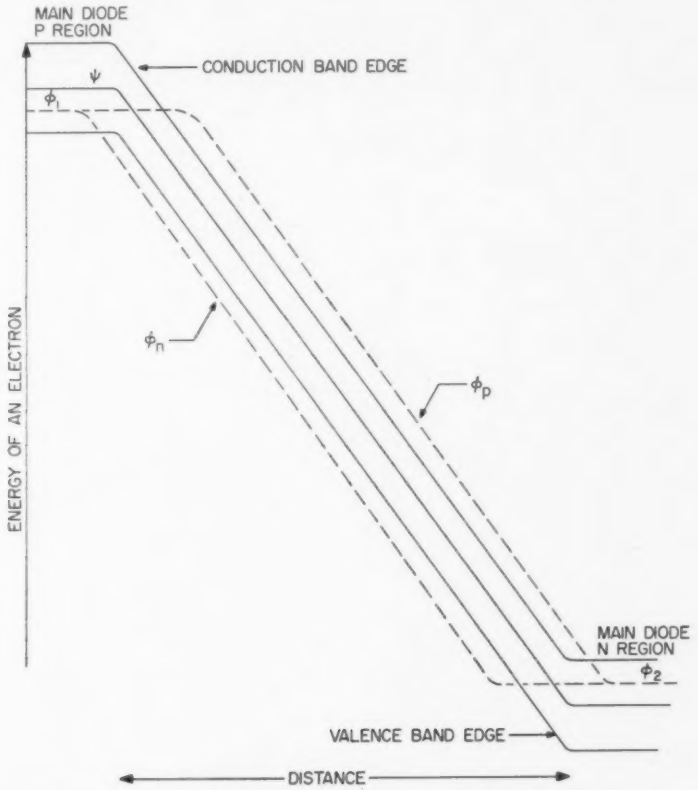


FIG. 9. Assumed band situation in the space-charge region.

where v_s is the saturated velocity ($v_s \simeq 6 \times 10^6$cm/sec for silicon at room temperature). Reverse-current densities may range from 10^{-4} A/cm^2 to 10^{-8} A/cm^2 suggesting that p may range from 10^8/cm^3 to 10^4/cm^3. Over this range of current density the hole density is much less than n_i. ϕ_p may lie as close as $4kT/q$ to the center of the gap or as far as $12kT/q$ from the center of the gap.

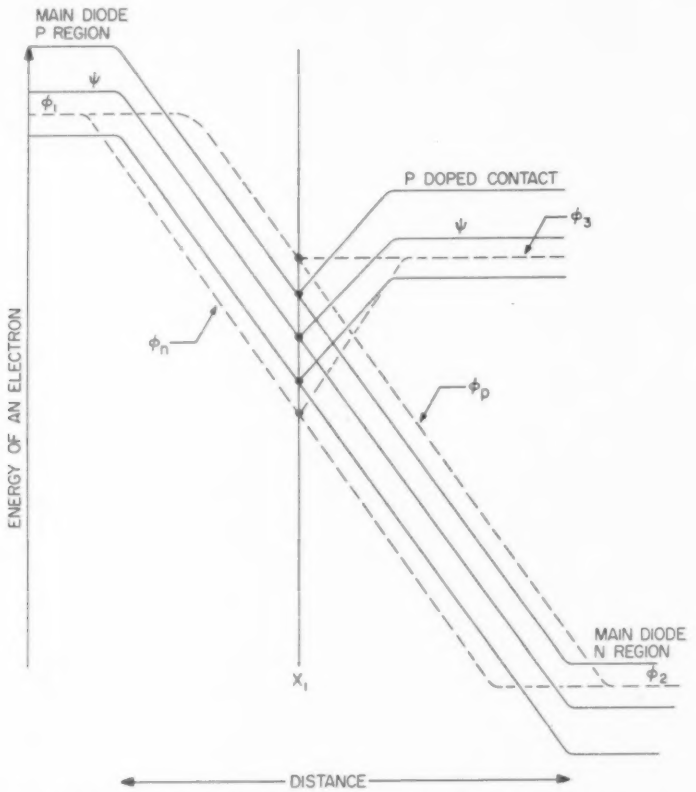


FIG. 10. Assumed band situation with a one-dimensional p-doped contact located in the space-charge region.

Similar arguments suggest that ϕ_n is located below the center of the gap. We require for our model that ϕ_n and ϕ_p be separated by a reasonable voltage, and this has been established.

Fig. 9 shows the band situation assumed to exist in the space-charge depletion region. For clarity we have again drawn the potential linear with distance and the quasi-Fermi levels outside the band gap. Far into the n region of the main diode we have a Fermi level characteristic of an electron-conducting semiconductor ϕ_2. Far into the p region of the main diode we have a Fermi level for a hole-conducting semiconductor ϕ_1. In the depletion region we have the non-equilibrium situation characterized by a quasi-Fermi level for electrons ϕ_n, and a quasi-Fermi level for holes ϕ_p. Since we are not primarily concerned with the edges of the depletion region, we have simply joined the Fermi levels and the quasi-Fermi levels with a smooth curve. The origin of potential, as is customary, is taken to be the center of the gap and denoted by ψ. It should be noted that the potential difference between the n and p regions is several hundred volts, and the distance between the two neutral regions is of the order of a few mils so that the bands are bent quite steeply. Fig. 10 shows the

band situation assumed to exist when a one-dimensional p-doped contact is alloyed into the space-charge region at the position x_1. In order to locate the p-doped contact in our energy-level diagram, we require that the Fermi level ϕ_3, far into the p-contact, join with the quasi-Fermi level for holes ϕ_p in the space-charge region at the point x_1. We then require that the band edges of the p-doped contact be joined to the band edges at the point x_1. Because an exact calculation depends upon many parameters and since the model does not require an exact knowledge of the shape of these curves, they have been drawn smoothly to the point of intersection x_1. We note that, between the p-doped contact and the p region of the main diode, a barrier to hole flow exists. In taking a hole from the p-doped contact we must first raise its potential before it can slide downhill in energy toward the p region of the main diode. It is this barrier which tends to keep the holes in the p-doped contact from flowing in the absence of an applied potential. Since the potential ψ changes sign in passing from the p-doped contact to the p region of the main diode, the negative gradient of ψ (the electric field) must go to zero between the two regions. Therefore, we do indeed have a situation in which $E = 0$ between the p-doped contact and the p region of the main diode and, therefore, a reasonable expectation that the current–voltage characteristic will be space-charge-limited.

(2) *Temperature-limited current flow.* If we return to Fig. 10 and examine the behavior of an electron in passing from the p-doped contact to the n–region of the main diode, we note that the electron flows continuously downhill in potential.* This suggests that no barrier to electron flow exists when a barrier to hole flow is present. That is, the injection of electrons from a p-doped contact will not be space-charge-limited. In order to inject electrons from a p-doped contact (in excess of the normal reverse-biased current) a concentration of electrons in excess of the equilibrium concentration is required. This can be obtained by superposing a small n-doped contact on the p-doped contact. Provided that the density of electrons in the n-doped contact is much greater than the density of holes in the p-doped contact, forward biasing of the small n–p structure will yield an injected excess density of electrons in the p-doped contact. Those electrons which appear at the interface between the p-doped contact and the space-charge region will be injected into the space-charge region and collected at the n side of the main diode.

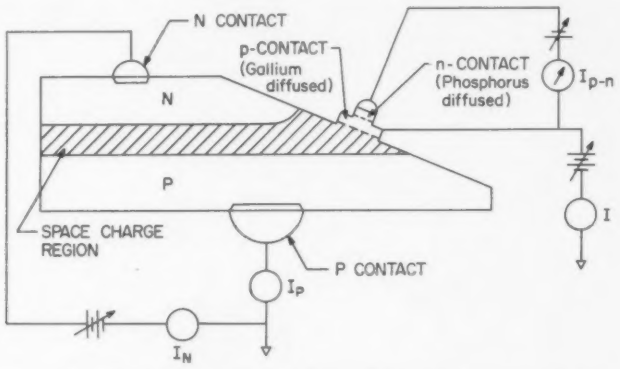


FIG. 11. The structure and bias used to obtain temperature-limited electron flow.

Fig. 11 shows the structure and the applied bias used to obtain injection of electrons from a p-doped contact into the space-charge region. The model given above suggests that the electron current injected from the p-doped contact will depend only upon the concentration of electrons at the p-contact-space-charge interface provided that the p-doped contact is reverse-biased for holes. Fig. 12 shows a plot of the injected electron current

* This is precisely the situation which occurs at the collector space-charge region of a junction transistor.

I_N as a function of I_{p-n}, the current flowing in the small p–n structure. The p-doped contact is biased at 40 V during this measurement. The data show that a large electron current can indeed be injected by this technique. The departure from linearity above 1·5 mA may be due to one or more reasons.

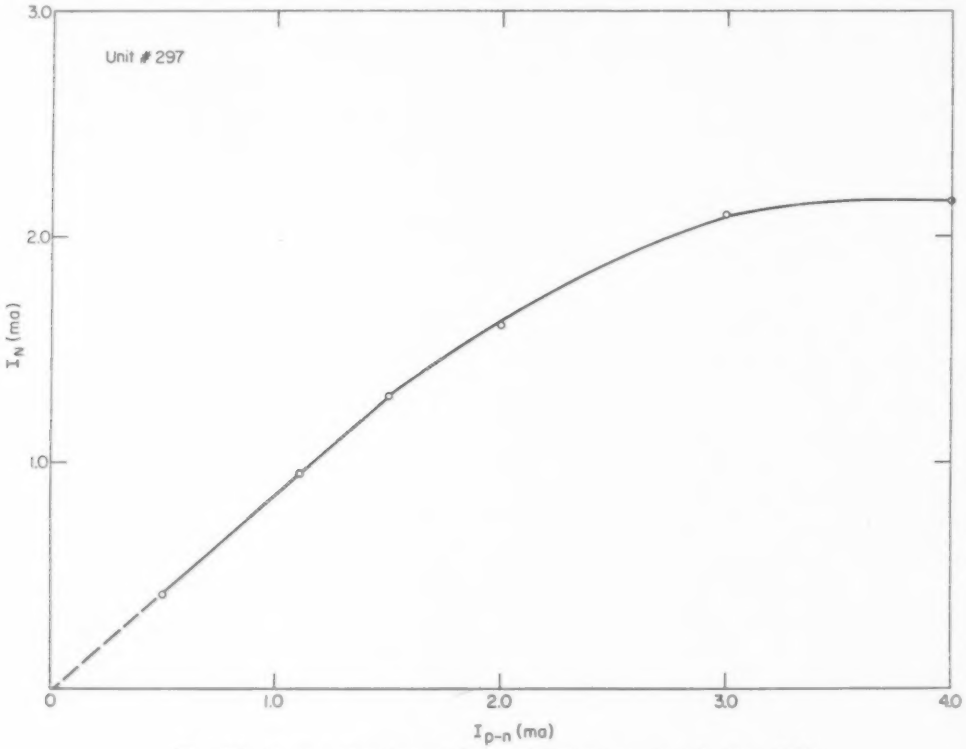


FIG. 12. I_N as a function of I_{p-n}, with the p-contact fixed at 40 V.

Firstly, the number of electrons reaching the p-contact–space-charge junction may fall off at high current densities owing to a reduction of injection efficiency of the n-contact. Secondly, the p-contact floating potential (as shown below) will vary with increasing electron density. The charge flowing out of the p-contact alters the potential in the space-charge region so that at an applied bias of 40 V the p-contact may no longer be reverse-biased for holes. Finally at very large current densities, space-charge effects, due to the current flow in the space-charge region, may limit the flow.

In order to observe the variation of the floating potential of the p-contact with increasing electron current, the small p–n diode was forward-biased and no contact was made between the p-contact and the n or p region of the main diode. Fig. 13 shows that the p-contact floating potential increases from about 40 V at $I_{p-n} = 0$ to about 60 V at $I_{p-n} = 1{\cdot}5$ mA. Therefore, at high I_{p-n} the p-contact is no longer reverse-biased for holes. This suggests that the saturation effect of I_N observed in Fig. 13 may be due to the erection of a barrier to electron flow when the barrier to hole flow is decreased. In Fig. 14 we show that an electron current I_N is injected even under p-contact floating conditions. In order for electrons to be injected when the p-contact is floating, an equal hole current must also be injected. Under these conditions, it is observed that a hole current I_P identically equal to I_N flows in the space-charge region.

The data of Figs. 13 and 14 indicate that with the p-contact floating, an electron current I_N dependent upon I_{p-n} is injected. In order to main-

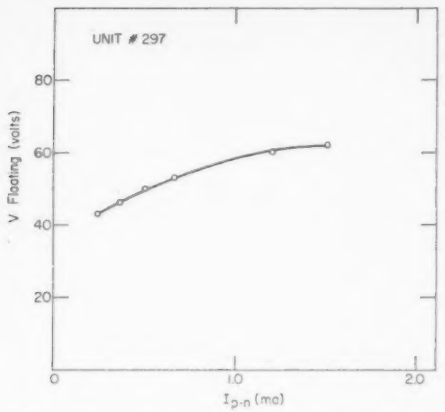


FIG. 13. p-Contact floating potential as a function of I_{p-n}.

tain no net current flow out of the p-contact an equal hole current must be injected, requiring that the potential of the p-contact increase.

From these floating potential measurements we may infer that the injected electrons may enhance the hole flow out of the p-contact. In Fig. 15 we show a plot of the hole flow I_P as a function of I_{p-n} with the potential of the p-contact fixed at 40 V. At low values of I_{p-n} we observe that the hole current is of the order of a few hundred microamperes. However, with increasing I_{p-n} (and with increasing I_N as shown in Fig. 12) the hole current

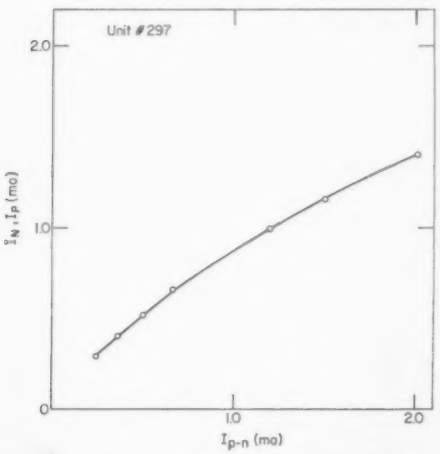


FIG. 14. I_N and I_P as a function of I_{p-n} with the p-contact floating.

increases to about 2 mA at $I_{p-n} = 4$ mA. Thus it appears possible to control the flow of holes out of the p-contact not only by varying the potential of the contact, but also by varying the injected electron density.

In Fig. 16 is shown a plot of the electron current I_N as a function of the p-contact voltage, with I_{p-n} as a parameter. We note that large electron current can be injected into the space-charge region by a p-doped contact, relatively independent of the voltage on the p-contact, provided that the contact is reverse-biased for hole flow. At high I_{p-n}, the electron current is dependent upon the p-contact voltage, but in this region the p-contact is injecting holes. This is seen more clearly in the data of Fig. 17 which shows the hole current as a function of the p-contact voltage with I_{p-n} as a parameter.

The injection of electrons from a p-doped contact which is reverse-biased for hole flow will be called temperature-limited. We have chosen this terminology by analogy with the vacuum diode case. In a vacuum diode, under temperature-limited conditions, every electron emitted by the cathode reaches the anode with no potential barrier present, in contrast to the space-charge-limited case in which a barrier to the flow of electrons exists between the cathode and the anode.

CONCLUSIONS

A plausibility argument has been presented to show that the injection of holes from a p-doped contact (and electrons from an n-doped contact) is probably space-charge-limited. On the other hand, the injection of electrons from a p-doped contact which is reverse-biased for hole flow (and holes from an n-doped contact which is reverse-biased for electron flow) is temperature-limited. In the former case a barrier to the carrier flow exists between the contact and the appropriate edge of the space-charge region; in the latter case no barrier to the flow exists. Experimentally, space-charge-limited injection restricts the injected current to low values even at relatively large voltage, while temperature-limited injection can yield large injected currents, relatively independent of the voltage applied to the contact.

Measurements of the floating potential of both p-doped and n-doped contacts have shown that an extremely high field region occurs at one edge of the contact, sufficiently large to cause avalanching

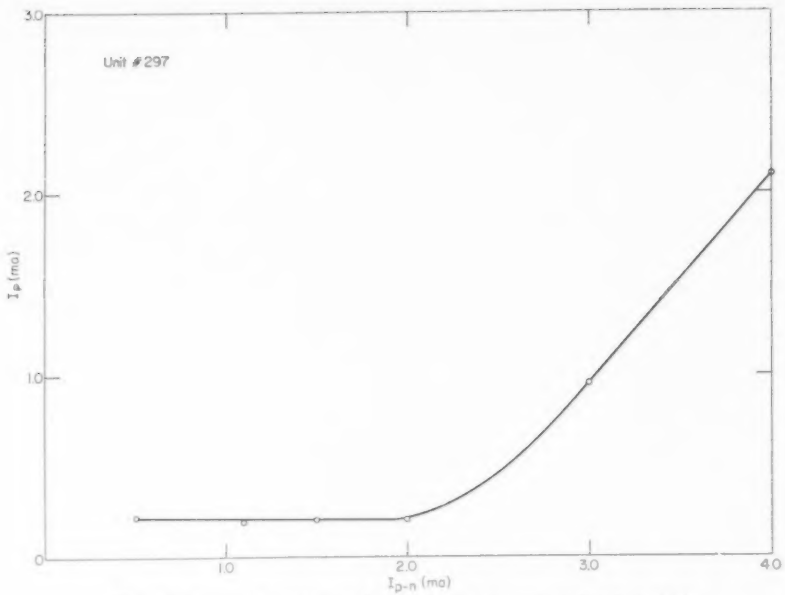


FIG. 15. I_P as a function of I_{p-n} with the p-contact fixed at 40 V.

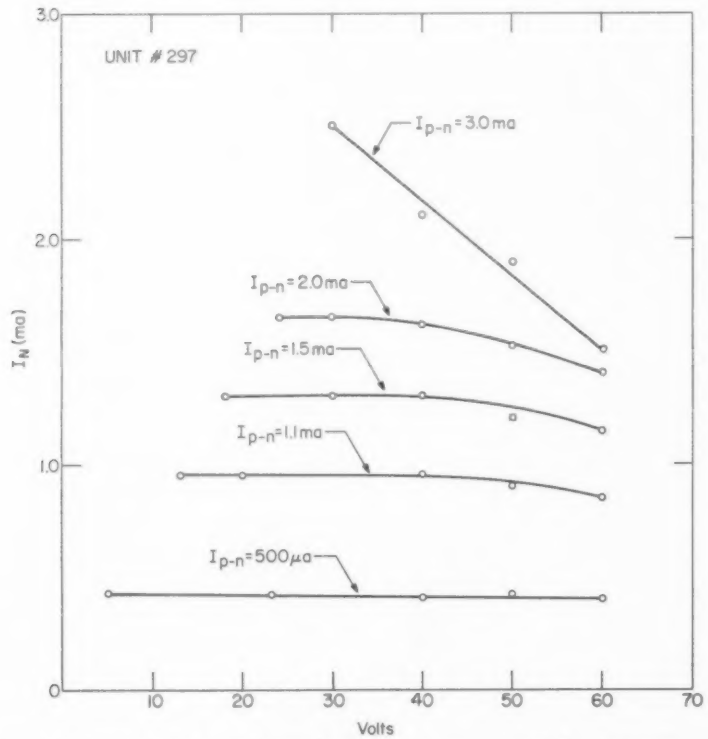


FIG. 16. I_N as a function of the voltage of the p-contact with I_{p-n} as a parameter.

at times, even in the absence of a potential applied to the contact.

It has been shown that under floating conditions temperature-limited injection of electrons occurs, which depends upon the excess density of electrons in the p-contact. The electron current results in an increase of the p-contact floating potential and an injected hole current identically equal to the electron current.

The presence of an enhanced hole flow from a p-doped contact in the presence of temperature-limited electrons has been shown. Hole current densities approaching those injected into neutral n-type material have been obtained from a p-doped contact in this manner. No model to explain this behavior has been presented, and we may only suggest here that the temperature-limited electrons are much more effective in reducing the barrier to hole flow than an applied voltage.

Finally, nothing has been stated about the mechanism of a point contact which can inject either electrons or holes. While it is possible to set up a simple model similar to that of Fig. 10, the behavior of a metal–semiconductor contact is never simple because of the non-intimate nature of the contact, particularly on silicon.

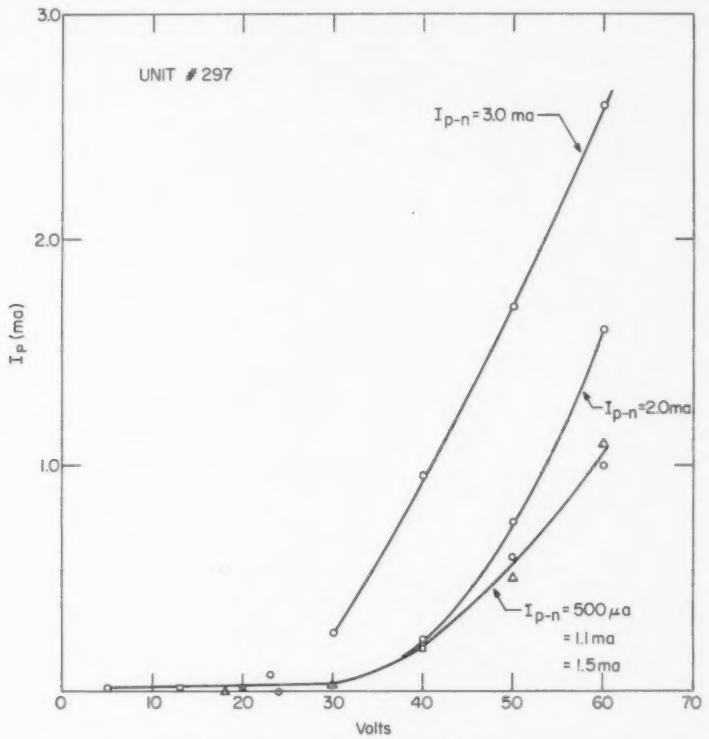


FIG. 17. I_P as a function of the voltage of the p-contact with I_{p-n} as a parameter.

Acknowledgements—The author would like to thank C. LANZA, A. GALEJS and R. NELSON for assistance in the fabrication and measurements on the units reported here. He would also like to thank the above individuals and Dr. H. STATZ for many informative discussions.

References

1. W. SHOCKLEY, *Bell Syst. Tech. J.* **28**, 435 (1949).
2. W. C. MATTHEI and F. A. BRAND, *J. Appl. Phys.* **28**, 513 (1957).

3. H. Statz, R. Pucel and C. Lanza, *Proc. I.R.E.* **45**, 1476 (1957).
4. W. Shockley and R. C. Prim, *Phys. Rev.* **90**, 753 (1953).
5. G. C. Dacey, *Phys. Rev.* **90**, 759 (1953).
6. E. J. Ryder, *Phys. Rev.* **90**, 766 (1953).
7. C. T. Sah, R. N. Noyce and W. Shockley, *Proc. I.R.E.* **45**, 1228 (1957).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 123–130. Printed in Great Britain

SUR LA DIFFUSION DES ÉLÉMENTS DES COLONNES III ET V DANS LE SILICIUM

R. SAINTESPRIT

Lignes Télégraphiques et Téléphoniques, Conflans-Ste. Honorine, France

(*Received* 27 *November* 1959)

Résumé—Ce papier concerne la diffusion de certains éléments des groupes III et V dans le silicium à la pression atmosphérique. Les trois éléments, gallium, bore et phosphore ont été étudiés en détail. Dés résultats quantitatifs sont donnés sur l'effet de protection d'une couche de silice, la témperature d'évaporation de l'impurité et la nature du gaz porteur. Les phénomènes d'oxydation et de réduction associés à la diffusion du gallium sont discutés, ainsi que l'affinité du bore pour l'oxyde de silicium.

Abstract—This paper deals with the diffusion of several elements from Groups III and V in silicon at atmospheric pressure. Gallium, boron and phosphorus have been considered in detail. Quantitative results are given on masking by a silicon dioxide layer, on the impurity evaporation temperature and the nature of the carrier gas. The oxidation and reduction phenomena associated with gallium diffusion are discussed, as well as the affinity of boron for silicon oxide.

Zusammenfassung—Diese Arbeit behandelt die Diffusion einiger Elemente der III. und V. Gruppe in Silizium bei Atmosphärendruck. Besonders Gallium, Bor und Phosphor werden betrachtet. Quantitative Ergebnisse werden angeführt über die Maskierung mit einer Siliziumdioxyd-Schicht, über die Verdampfungstemperatur der Verunreinigungen und über die Natur des Trägergases. Die in Verbindung mit der Gallium-Diffusion auftretenden Oxydations- und Reduktions-Phänomene werden besprochen, ebenso die Affinität des Bors zum Siliziumoxyd.

CETTE note concerne les diffusions vapeur–solide d'impuretés dans le silicium sous une pression voisine de la pression atmosphérique, conformément à la technique de FROSCH[1].

Ces diffusions sont, en fait, utilisées pour former des structures de transistors: elles doivent donc permettre de réaliser, de façon reproductible, des couches contaminées par des impuretés diffusées, d'épaisseurs et de concentrations définies. En outre, la disposition de ces couches doit être telle que l'on puisse y déposer des contacts et avoir un accès électrique à chacune d'elles. La Fig. 1 donne le schéma de la structure envisagée. Le matériau dont on part est du type *n*: il constitue le collecteur du transistor; une couche *p*, créée par diffusion, forme la base du transistor; dans une partie de cette couche *p*, une couche diffusée de type *n*, à forte concentration, forme l'émetteur du transistor. Les parties de la couche *p* dans lesquelles l'émetteur ne doit pas diffuser sont protégées par une oxydation du silicium. De telles structures ont déjà été décrites, en particulier par ASCHNER *et al.*[2]

Lors des diffusions, les phénomènes fondamentaux résident dans les réactions entre la surface du silicium et le milieu extérieur. Ces réactions commandent la concentration superficielle des impuretés diffusées et la loi de variation de la concentration de ces impuretés à l'intérieur du silicium: l'effort principal porta donc sur l'étude de ces réactions.

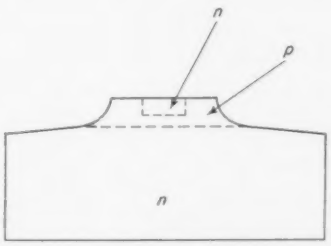


FIG. 1.

Du point de vue théorique, le problème est complexe: l'expérimentation s'est donc pratiquement d'abord effectuée sans ordre préétabli, avant d'arriver aux quelques hypothèses qui seront indiquées plus loin, et dont le principal intérêt fut de guider la suite de l'étude.

1. MODE OPERATOIRE

Ce premier chapître résume le mode opératoire, classique à présent, adopté pour cette étude. Dans cette méthode, purement dynamique, un courant gazeux, chargé d'un certain pourcentage d'impureté, lèche le silicium. Ce pourcentage d'impureté dépend de la température de la source, de la nature, c'est-à-dire du poids et des dimensions, de la molécule de l'impureté, de la nature du gaz d'entraînement, de son débit, et de la loi de répartition de la température le long de l'axe du four. L'évaporation du silicium, de la silice éventuellement formée, des impuretés présentes dans le silicium, soulèvent également des problèmes, dynamiques comme le précédent.

Les traitements thermiques nécessaires à la diffusion présentent l'inconvenient de diminuer la durée de vie des porteurs minoritaires dans le silicium. Toutefois, des vitesses de refroidissement relativement lentes évitent en partie cet inconvénient. D'autres dispositions sont également prévues: elles font actuellement l'objet d'essais.

L'appareillage est simple: il comprend un four dit à deux ou trois températures. La Fig. 2 montre le schéma d'un tel four.

Le tube du four t et ses embouts e_1, e_2 sont en quartz soufflé.

L'impureté est placée dans une nacelle en quartz soudée à la gaîne du thermocouple th 1; ceci permet de contrôler la température d'évaporation des impuretés que l'on porte pour cela dans les zones θ_1 ou θ_2 du four. Lorsque la source d'impuretés est de l'anhydride borique, on interpose une nacelle d'alumine.

Les plaquettes de silicium sont posées dans les fentes d'un support en quartz, soudé à la gaîne du thermocouple th 2. Ces plaquettes sont chauffées dans la région θ_3 du four pour la diffusion, et peuvent être chauffées dans la région θ_2, lors d'une contamination préalable, à une température inférieure à la température de diffusion.

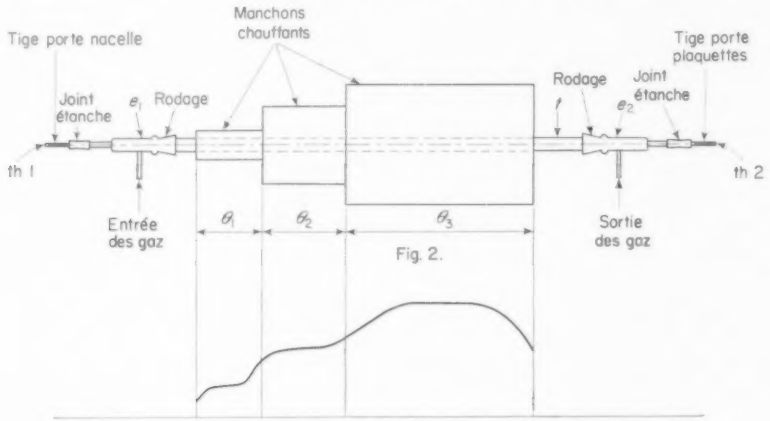


Fig. 2.

FIGS. 2 et 3.

Le calorifugeage et l'écartement des zones θ_1, θ_2 et θ_3 sont tels qu'à l'intérieur du tube t, sur son axe, à l'équilibre thermique, le diagramme des températures peut se représenter par la courbe de la Fig. 3.

Dans les zones θ_1 et θ_2, la température varie lentement (moins de 10°/cm): dans la zone θ_3, la température est constante à moins de 5°C près dans une région de 10 cm de longueur environ. Un thermocouple placé près des enroulements de chauffage, à l'extérieur du tube de quartz, régule la température de la zone θ_3 à moins de 2°C près.

Cet équipement permet de réaliser un programme quelconque de montée en température des plaquettes de silicium et des impuretés, dans

l'atmosphère désirée, ainsi qu'un refroidissement aussi lent qu'il est nécessaire.

2. METHODES DE MESURES UTILISEES

2.1 *Epaisseur de la couche*

La profondeur de la jonction est mesurée au microscope métallographique, après avoir taillé un biseau de 5° dans la plaquette de silicium, et coloré les zones de type *p* en brun par l'application d'une goutte d'acide fluorhydrique, sous un éclairage assez fort.

2.2 *Concentration superficielle de la couche*

La concentration superficielle C_0 se calcule à partir de la résistance de la couche par cm^2 de surface, compte tenu de la loi de variation de la concentration en fonction de la pénétration. Ce calcul est basé sur les hypothèses suivantes:

(1) l'évaporation du silicium et de la silice, à la pression atmosphérique, est négligeable;

(2) la contamination du silicium se fait de l'une ou l'autre des deux manières suivantes:

(a) diffusion à concentration constante: l'impureté s'évapore pendant tout le temps de la diffusion; la concentration, à la distance x de la surface, est donnée par l'expression:

$$C_x = C_0 \operatorname{erfc} \frac{x}{2\sqrt{(Dt)}}$$

(b) diffusion à quantité constante: l'impurité s'évapore pendant un temps négligeable devant le temps de la diffusion; les plaquettes de silicium sont portées à une température faible pendant l'évaporation de l'impureté, de telle sorte qu'à la fin de cette évaporation, la contamination du silicium ne soit sensible qu'à une profondeur négligeable devant la profondeur de la jonction finale. La loi de variation de la concentration en fonction de la pénétration x est, dans ce cas:

$$C_x = C_0 \exp(-x^2/4Dt)$$

La loi de variation de la concentration de l'impureté diffusée dans le silicium peut être vérifiée expérimentalement: au moyen de la solution d'attaque chimique lente[3], on élimine, approximativement micron par micron, les couches externes de la plaquette de silicium, et l'on mesure chaque fois la résistance en surface; la comparaison des valeurs de résistance ainsi obtenues et de l'intégrale de la concentration, donnée par l'une ou l'autre des formules précédentes, permet de définir la loi de variation de la concentration dans le silicium. Ces mesures, assez longues à effectuer, n'ont été faites que sur certaines jonctions, d'un intérêt particulier.

2.3 *Durée de vie des porteurs libres*

La méthode utilisée est celle de la décroissance de la photoconductivité. La mesure est faite sur des plaquettes de silicium ayant aubi un traitement thermique identique à celui des diffusions, mais sans contamination superficielle, afin d'éviter la formation des jonctions qui gêneraient la mesure.

3. RESULTATS OBTENUS

Ces résultats concernent essentiellement les diffusions de gallium, de bore, de phosphore et d'arsenic.

3.1 *Diffusion de gallium*

3.1.1. *Source d'impuretés: sesquioxyde de gallium* Ga_2O_3. Le sesquioxyde de gallium est très stable au dessous de 1300°.[4] A 1000°C sa tension de vapeur n'est pas négligeable. A cette température, nous avons évaporé Ga_2O_3 sous argon. Il se condense à la sortie du four, sans transformation apparente, et sans avoir contaminé le silicium.

Il est possible de contaminer le silicium par le sous-oxyde de gallium: Ga_2O, qui est instable et se sublime au-dessus de 500°C. Ce monoxyde est formé directement dans le tube de diffusion, par exemple en chauffant Ga_2O_3 sous hydrogène humide.[1] Lorsque Ga_2O_3 atteint 600°C environ, une brusque réaction se produit. Le contenu de la nacelle devient noir instantanément, vraisemblablement par le gallium pulvérulent libéré, alors qu'à la sortie du four, différents produits de condensation appraissent. Dans une région où la température est voisine de 500°C se condense un produit brun, qui est vraisemblement du sous-oxyde de gallium Ga_2O. Dans une région plus froide, le gallium métallique se dépose.

Si le sesquioxyde de gallium partiellement décomposé est maintenu vers 600°, la contamination du silicium apparaît très faible, on obtient sur le silicium une mince pellicule irrégulière de type *p*. Si le sesquioxyde de gallium est monté vers 900°, le silicium paraît contaminé avec une concentration en surface qui peut atteindre 10^{17}

at/cm³ pour un temps de diffusion court (moins de 2 hr). Pour des temps de diffusion plus longs (3–5 hr), la concentration en surface paraît décroître.

Pour les diffusions longues ($\approx$5 hr), une mince couche n apparaît sur la couche p en confirmation de résultats déjà publiés.[5]

D'après GASTINGER (informations personnelles), la réaction qui se produit aux environs de 600° serait la suivante:

$$Ga_2O_3 + 2H_2 \rightarrow Ga_2O + 2H_2O \qquad (1)$$

$$Ga_2O + H_2 \rightarrow 2Ga + H_2O \qquad (2)$$

Le sous-oxyde Ga_2O se sublime au dessus de 500°.

3.1.2. *Source d'impuretés: gallium métal.* La réaction (3) nous a conduit à travailler avec du gallium pur, en atmosphère d'hydrogène humide (barbottant dans l'eau à 18°C). Le gallium est chauffé entre 800 et 900°C. Il fond et s'entoure d'oxyde comme le ferait l'aluminium. Dans l'hydrogène humide, à ces températures, doit se former assez lentement le monoxyde Ga_2O qui réagit à chaud avec le silicium [formule (4)].

Le Tableau 1 donne quelques résultats. La résistivité du silicium n de base variait entre 4 et 10Ω; la concentration était donc négligeable devant la concentration superficielle de la couche p diffusée.

Quelle que soit la valeur de la concentration C_0, elle est uniforme sur toute la surface des plaquettes

Tableau 1.

Température des plaquettes de Si (°C)	Température du Ga (°C)	Durée de la diffusion (min)	Concentration superficielle (at/cm³)	Résistance de la couche (Ω/cm²)
1220	850	30	2×10^{17}	850
1220	860	30	$3{,}5 \times 10^{17}$	500
1220	870	30	$5{,}5 \times 10^{17}$	330
1220	900	30	$1{,}2 \times 10^{18}$	200
1220	950	30	$1{,}2 \times 10^{18}$	200

Ceci explique l'apparition brutale, à la sortie du four, des produits de décomposition du sous-oxyde. Cette évaporation brutale du sous-oxyde n'est pas susceptible de contaminer suffisamment le silicium, à cause peut-être de la rapidité de la réaction. Par contre, il semble que le gallium produit dans la nacelle peut, lorsqu'il est porté à plus haute température, au-delà de 800°, former plus lentement du sous-oxyde.

$$2\,Ga + H_2O = Ga_2O + H_2 \qquad (3)$$

La contamination du silicium se ferait alors par réaction du sous-oxyde avec le silicium.

$$2\,Ga_2O + Si = 4\,Ga + SiO_2 \qquad (4)$$

Dans nos expériences, le sous-oxyde de gallium, brutalement libéré [formule (1)] était projeté dans le tube de quartz et formait des silicates friables. Ce tube devenait très cassant et perdait en particulier sa résistance aux chocs thermiques.

de silicium, dont le diamètre moyen était de 20 mm.

3.1.3. *Action de la silice.* Dans ce procédé de diffusion de gallium, à concentration constante, le silicium se recouvre d'une couche de silice épaisse de plusieurs milliers d'angströms. Jusqu'à 1250°C, l'oxyde ne paraît avoir aucune action sur la contamination du silicium par le gallium. Cet oxyde est utilisé dans la fabrication des structures de transistor. Il est indifférent de le former avant ou après la contamination du silicium par le gallium.

En contrepartie, lors de la diffusion à quantité constante, la concentration superficielle diminue très fortement en surface, sans doute à cause de l'évaporation du gallium au cours de la diffusion. Pour éviter cet appauvrissement en surface, on peut évaporer en fin de diffusion une petite quantité de gallium qui, augmentant la concentration de surface C_0, facilite la formation d'un contact ohmique.

3.2 *Diffusion de bore.*

La source d'impureté est de l'anhydride borique. La contamination du silicium dans ces conditions est assez délicate. Sans précaution, la concentration superficielle n'est pas constante sur la plaquette de silicium, et les essais ne sont pas reproductibles.

Il semble que ces difficultés proviennent de deux causes:

(a) l'anhydride borique fixe l'eau très facilement en donnant BO_3H_3; ce corps est entraînable par l'eau. La tension de vapeur varie fortement lorsque l'anhydride borique est hydraté.[6]

(b) l'anhydride borique est très stable.[7] Il n'est pas rèduit par le silicium à chaud.[8] La silice est soluble dans B_2O_3 fondu.[9]

Ceci peut expliquer les premiers résultats obtenus.

Pour obtenir des résultats reproductibles, il faut déshydrater l'acide borique pendant au moins 2 hr à une température supérieure à 1000°C. La concentration superficielle n'est uniforme sur les plaquettes de silicium qu'à la condition de les maintenir dans une atmosphère gazeuse à concentration constante de bore; c'est-à-dire sous un gaz neutre à très faible débit. Il est difficile d'expliquer la contamination du silicium par l'anhydride borique. On doit être en présence d'une solubilité de la vapeur B_2O_3 dans le silicium, voisine de celle d'un gaz dans un métal.

Le Tableau 2 donne quelques résultats obtenus en atmosphère d'azote sec sur une plaque de silicium non oxydée (épaisseur d'oxyde non mesurable).

Tableau 2.

Température des plaquettes de Si (°C)	Température de B_2O_3 (°C)	Durée de l'évaporation (min)	Durée de la diffusion (hr)	Concentration superficielle (at/cm³)	Résistance de la courbe (Ω/cm²)
1220	1060	15	2	3×10^{18}	60
1220	1085	15	2	$1{,}2 \times 10^{19}$	25
1220	1120	30	2	5×10^{19}	9
1220	1200	30	3	6×10^{20}	1,8

Action de la silice:

Une faible épaisseur d'oxyde gêne la contamination du silicium par les vapeurs d'anhydride borique.

Le Tableau 3 montre l'action dans le temps d'une couche de silice faible formée sur le silicium avant contamination.

D'après les indications de ce tableau, on voit que, à la température de diffusion à laquelle nous travaillons, que l'on peut difficilement dépasser sans risquer le ramollissement du tube de quartz, il n'est pas possible de faire diffuser du bore après avoir oxydé le silicium: la diffusion du bore doit donc être faite en atmosphère non oxydante. Bien que les plaquettes de silicium, après une telle diffusion, soient toujours recouvertes d'une très légère couche d'oxyde, cette couche est néanmoins trop mince (épaisseur non mesurable) et très insuffisante pour assurer la protection contre la contamination par l'impureté de type *n*, au cours de la deuxième diffusion formant l'émetteur.

Tableau 3.

Température des plaquettes de Si (°C)	Épaisseur d'oxyde (Å)	Température de B_2O_3(°C)	Durée de l'évaporation (min)	Durée de la diffusion (hr)	C_0 (at/cm³)	Profondeur de la jonction (μ)
1220	1000	1200	20	2	—	—
1220	1000	1200	20	19	10^{18}	20

Il faut donc procéder à l'oxydation du silicium après la diffusion du bore; mais cette opération risque également de présenter un inconvénient: le coefficient de partage du bore entre le silicium et la silice est vraisemblablement très inferieur à l'unité, puisque la concentration superficielle du bore dans le silicium, trouvée sous la couche d'oxyde après une diffusion de 19 hr, est faible vis-à-vis des concentrations obtenues dans du silicium non oxydé; c'est probablement la raison pour laquelle l'oxydation du silicium après diffusion du bore réduit très fortement la concentration superficielle du bore dans le silicium; tout se passe comme si, lors de l'oxydation, le bore se partageait entre le silicium et la couche de silice superficielle pour laquelle il a une affinité.

Pour éliminer cet inconvénient, il suffit d'effectuer l'oxydation superficielle du silicium en présence d'anhydride borique, de façon telle que l'oxyde formé se sature de bore à partir des vapeurs d'anhydride borique et non pas à partir du bore contenu dans le silicium.[10]

Le Tableau 4 montre l'action de l'oxydation sur une plaque recouverte d'une couche p due au bore, d'une part en l'absence de vapeur d'anhydride borique, d'autre part en présence de vapeur d'anhydride borique.

Tableau 4.

Température des plaquettes de Si (°C)	Température de B_2O_3 (°C)	Conditions d'oxydations ultérieures	Nombre d'atomes diffusés dans la couche (at/cm^2)
1220	1075	—	5×10^{15}
1220	1075	1 hr sous O_2	2×10^{14}
1220	1075	1 hr sous O_2 + vapeur B_2O_3	3×10^{15}

3.3. *Diffusion de phosphore*

3.3.1. *Source d'impuretés: anhydride phosphorique.* L'anhydride phosphorique est décomposé à chaud par le silicium. Cette réaction permet, comme pour le gallium, une contamination facile du silicium, et la formation de couches de concentration superficielles uniformes et reproductibles. Aux températures de diffusion, la solubilité limite du phosphore dans le silicium est élevée. Pour diminuer d'une façon appréciable la concentration, il faut travailler avec des tensions de vapeur si faibles qu'il devient alors difficile, par cette technique de tube ouvert, d'obtenir des couches uniformes et reproductibles. Mieux vaut réduire la température du silicium au cours d'une prédéposition. Ce procédé a montré l'avantage suivant: lorsque la contamination du silicium se fait à température relativement élevée (entre 1000 et 1100°C), il se forme très rapidement une couche à forte concentration. Tout se passe comme si la constante de diffusion augmentait lorsque la concentration croît. Ce phénomène est très gênant

Tableau 5.

Températures du Si pour la prédéposition (°C)	Température de P_2O_5 (°C)	Durée de la prédéposition (min)	Températures du Si pour la diffusion (°C)	Durée de la diffusion (hr)	Résistance de la couche (Ω/cm^2)	Concentration superficielle (at /cm^3)
950	260	30	1150	1	2,5	1×10^{21}
1000	260	30	1150	1	1,4	2×10^{21}
1200	260	15	1200	1	1,5	$1,5\times10^{21}$

Tableau 6.

Température du Si (°C)	Température de PCl^5 (°C)	Durée de la diffusion (hr)	Résistance de la couche (Ω/cm^2)	Concentration superficielle (at/cm^3)
1125	120	1	1,8	3×10^{21}

lorsque l'on veut obtenir des couches minces. Il faut, dans ce cas, faire une predéposition de phosphore sur des plaques portées à une température inférieurs à 1000°C, et les monter ensuite à une température compatible avec la profondeur et le temps de diffusion désirés.

3.3.2. *Source d'impuretés: pentachlorure de Phosphore.* Pour éviter la formation de cette couche superficielle à forte concentration, des essais ont été faits avec des composés non oxygénés du phosphore. Les résultats n'ont pas été satisfaisants. Le Tableau 6 donne un résultat obtenu à partir du pentachlorure de phosphore; la couche formée présente la même croûte à forte concentration que celle que l'on obtient à partir du pentoxyde de phosphore.

3.3.3. *Action de la silice.* Aux températures utilisées, comprises entre 1050 et 1100°C, le phosphore se prête bien à la protection du silicium par l'oxyde. Il n'en est pas de même aux températures plus élevées.

Le Tableau 7 montre cette action.

Tableau 7.

Température du Si(°C)	Epaisseur de l'oxyde (Å)	Température de P_2O_5 (°C)	Durée de l'évaporation (min)	Durée de la diffusion (hr)	Résistance de la couche (Ω/cm^2)	Concentration superficielles (at/cm^3)
1100	3000	220	10	2	—	—
1120	3000	220	10	4	>1500	$<10^{16}$
1200	3000	220	10	4	0,7	$1,3 \times 10^{21}$

3.4. *Diffusion d'arsenic*

La source d'impureté est l'anhydride arsénieux; comme l'anhydride phosphorique, l'anhydride arsénieux est décomposé à chaud par le silicium. La contamination du silicium ne présente pas de difficulté: les couches contaminées obtenues ont des concentrations régulières. L'inconvénient de l'arsenic vient de la faible valeur de sa constante de diffusion dans le silicium; elle oblige, pour contaminer la zone formant l'émetteur du transistor, à faire diffuser l'arsenic à des températures et pendant des temps tels qu'une couche d'oxyde, épaisse de quelques 0,1μ, ne protège plus, contre

Tableau 8.

Température du Si(°C)	Epaisseur de l'oxyde (Å)	Température de As_2O_3 (°C)	Durée de l'évaporation (hr)	Durée de la diffusion (hr)	Résistance de la couche (Ω/cm^2)	Concentration superficielle (at/cm^3)
1200	0	220	3	3	120	4×10^{18}
1250	0	220	5	5	55	4×10^{18}
1180	3000	220	3	3	—	—
1180	3000	220	3	6	1200	7×10^{16}

la contamination, les parties de la surface du silicium qui doivent rester du type de résistivité p de la base.

Le Tableau 8 donne quelques-uns des résultats obtenus.

4. CONCLUSIONS: DIFFUSIONS POUR UNE STRUCTURE DE TRANSISTOR

Les travaux que l'on vient d'exposer ont conduit à adopter le gallium et le phosphore pour réaliser, dans les meilleures conditions, la structure du transistor désiré.

Les plaquettes de silicium sont oxydées, puis leur surface est contaminée au gallium, à partir de gallium pur, en une même opération thermique, suivie d'un refroidissement très lent. Aux emplacements réservés aux émetteurs, la couche de silice est enlevée par un procédé de photogravure. Les émetteurs sont alors formés par une diffusion sélective de phosphore, à partir d'anhydride phosphorique. Les épaisseurs de la base et de l'émetteur sont respectivement de l'ordre de 2 et 4 μ.

Remerciements—Ces travaux ont été effectués au titre d'un contrat entre le Service Technique des Télécommunications de l'Air et la Société L.T.T. A l'un et à l'autre, l'auteur adresse ses remerciemente pour l'avoir autorisée à publier ses résultats.

BIBLIOGRAPHIE

1. C. J. Frosch et L. Derick, *J. Electrochem. Soc.* **104**, 547 (1957).
2. J. F. Aschner, C. A. Bittmann, W. F. J. Hare et J. J. Kleimack, *J. Electrochem. Soc.* **106**, 415 (1959).
3. J. Andrus et W. L. Bond, *Transistor Technology* (F. J. Biondi, Rédacteur) Vol. III, p. 158. Van Nostrand, London (1958).
4. *Gmelin's Handbuch der Anorganischen Chemie* No. 36, Gallium, p. 69. Verlag Chemie, Berlin (1936).
5. C. J. Frosch et L. Derick, *J. Electrochem. Soc.* **105**, 695 (1958).
6. *Gmelin's Handbuch der Anorganischen Chemie* No. 13, Bor, p. 140. Verlag Chemie, Weinheim (1954).
7. C. Brunold, *Chimie Général et Métalloïdes* p. 681. Masson et Cie., Paris (1952).
8. P. Pascal, *Traité de Chimie Minérale* Tome V, Silicium, p. 381. Masson et Cie., Paris (1932).
9. *Gmelin's Handbuch der Anorganischen Chemie* No. 13, Bor, p. 135. Verlag Chemie, Weinheim (1954).
10. R. Saintesprit. Proposition de brevet 776.083 du 7 Octobre, 1958.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 131–134. Printed in Great Britain

INDIUM AS A SOURCE OF IMPURITIES IN INDIUM ARSENIDE*

A. J. STRAUSS, T. C. HARMAN, E. B. OWENS and M. C. FINN

Lincoln Laboratory, Massachusetts Institute of Technology, Lexington 73, Massachusetts

(*Received* 17 *November* 1959; *in revised form* 14 *December* 1959)

Abstract—The sulfur content of high-purity indium samples from five commercial suppliers, as determined by a colorimetric method, has been found to range from 0·36 p.p.m. (by weight) to less than 0·05 p.p.m. (by weight), the limit of detection. These results, together with recent data on the purification of arsenic by treatment with molten lead, show that a significant proportion of the donor impurities in the purest InAs now available may originate in the indium used for synthesis.

Résumé—La teneur en soufre d'échantillons d'indium très pur se trouve être entre 0,36 p.p.m. et 0,05 p.p.m. par poids qui est la limite de détection des méthodes calorimétriques. Ces résultats avec les données récentes sur la purification de l'arsenic par traitement avec du plomb fondu montre qu'une partie notable des impuretés donneurs dans l'InAs le plus pur peut provenir de l'indium utilisé pour la synthèse.

Zusammenfassung—Der Schwefelgehalt hochreiner Indiumproben, die von fünf verschiedenen Lieferfirmen stammten, wurde kolorimetrisch bestimmt; es wurden Werte (in Gewichsteilen) von 0,36 p.p.m. bis zu 0,05 p.p.m., der Nachweisgrenze der Methode, ermittelt. Diese Ergebnisse, zusammen mit kürzlich ermittelten Daten über die Arsen-Reinigung durch Behandlung mit geschmolzenem Blei, lassen erkennen, dass ein bedeutender Anteil der Donatoren im reinsten InAs, das z.Z. erhältlich ist, aus dem für die Herstellung verwendeten Indium herrührt.

INTRODUCTION

EFFORTS made at a number of laboratories to prepare pure indium arsenide have been only moderately successful. The purest InAs reported in the literature[1,2] is estimated to contain donor impurity concentrations of approximately 1–2 × 10^{16} cm^{-3}—orders of magnitude greater than the concentrations of electrically active impurities in the purest germanium, silicon and indium antimonide. No significant reduction in impurity concentration has been reported since 1956.

In the absence of analytical data establishing the identity and origin of the donors in the purest InAs now available, it has generally been assumed that the donors are sulfur and perhaps also selenium atoms. These elements are reported to have distribution coefficients in InAs of 1·0 and 0·93, respectively,[3] which are consistent with the slowly varying free-electron concentrations observed in zone-refined ingots. It has also been generally assumed that the donors originate primarily in the arsenic used in synthesis. This assumption has been justified chiefly on the ground that previous improvements in the purity of InAs have been achieved by using arsenic treated in a manner expected to remove sulfur and selenium (for example, treatment with molten lead or sublimation in vacuum or hydrogen). In addition, it has been argued that the donor concentrations in the high-purity indium used to prepare InAs must be much less than 10^{16} cm^{-3}, since InSb with total impurity concentrations of the order of 10^{14} cm^{-3} can be prepared from the same indium.

When the foregoing arguments are examined carefully, they are found to be invalid. The fact that the purity of InAs has been increased by using specially treated arsenic shows that untreated arsenic contains relatively high donor concentrations, but it does not demonstrate that the donor concentrations are still significant in the purified

* The work reported in this paper was performed by Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology with the joint support of the U.S. Army, Navy and Air Force.

arsenic. On the other hand, since high-purity InSb is prepared by extensive zone refining, impurities with low distribution coefficients in InSb could be present in indium at concentrations much greater than the total impurity concentration in the purest InSb made from the indium. This could certainly be the case for sulfur and even to some extent for selenium, whose effective distribution coefficients in highly doped zone-refined ingots of InSb have been estimated to be 0·1 and 0·5, respectively.[4] (Measurements on pulled single crystals of InSb show that the distribution coefficient of selenium decreases with decreasing selenium concentrations.[5,6] Therefore, the effective distribution coefficient of selenium in high-purity zone-refined InSb ingots may be significantly less than the value of 0·5 estimated for highly doped ingots.)

In order to investigate the role of indium as a source of impurities in InAs, the sulfur content of high-purity indium from five commercial suppliers has been determined by chemical analysis. The analytical results show that a number of these samples would contribute sulfur concentrations of the order of 10^{16} cm^{-3} to InAs prepared from them. Therefore, it seems probable that significant proportions of the impurities in some samples of high-purity InAs do originate in the indium. This finding is supported by recent results which show that arsenic purified by treatment with molten lead may contribute sulfur and selenium concentrations of less than 10^{15} cm^{-3} to InAs.[7]

ANALYTICAL METHOD

The sulfur determinations were made by the method of LUKE,[8] who reported its application to a number of metals other than indium. In this method, sulfur is oxidized to the sulfate during sample dissolution, reduced to the sulfide, and distilled as hydrogen sulfide into ammonium hydroxide solution. Lead citrate is added and the concentration of lead sulfide determined colorimetrically.

In order to obtain higher sensitivity than LUKE reported, the sample size was increased, and a novel sampling procedure was employed to reduce uncertainties due to the relatively high reagent blanks encountered. In this procedure, the sulfur content of an indium sample is not compared with that of a reagent blank determined in the absence of indium; instead, sulfur determinations are carried out on 0·5-g and 5·5-g samples of indium, and also on a 5·5-g indium sample to which 1 μg of sulfur is added. Since the quantity of reagents used is the same for all three samples, the difference in total sulfur between the 0·5-g sample and the 5·5-g sample without added sulfur can be taken as the sulfur content of 5 g of indium; the method is standardized by comparing the readings for the two 5·5-g samples with and without added sulfur. By use of this procedure, the limit of detection for sulfur in indium has been reduced to 0·05 p.p.m. (by weight). A standard deviation of 0·05 p.p.m. (by weight) was obtained in measurements on fifteen 5-g samples of indium.

The details of the technique which was employed in applying LUKE's method to indium are as follows: The analysis was carried out on duplicate portions of the etched indium sample, each weighing 12·5 g. Each portion was dissolved in 50 ml of 1 : 1 HCl : HNO_3 mixture. The resulting solution was reduced in volume by evaporation and then transferred quantitatively to a 50-ml volumetric flask. The washing for transferring was done with HCl. From the solution an aliquot equivalent to 0·5 g of indium and two aliquots each equivalent to 5·5 g of indium were pipetted into reaction flasks. An aliquot solution containing 1·00 μg of sulfur as K_2SO_4 was added to one of the 5·5-g aliquots. Each flask was then heated to reduce the volume of solution to about 2 ml. The walls of the flask were rinsed down with 10 ml of HCl, and the volume again reduced by evaporation to about 2 ml. This reduction in volume and rinsing with HCl was repeated until no yellow coloration could be seen in the solution. The last traces of HNO_3 were removed by slowly adding successive 0·2-ml portions of formic acid, while the solution was being heated on a hot plate with surface temperature of 135°C, until no brown fumes were seen. The solution was then placed on a hot plate with surface temperature of 185°C and boiled to a volume of 2–3 ml. Next, 10 ml of HCl were added, and volume reduced by evaporation to 5 ml to remove the last traces of excess formic acid. Then 25 ml of 4:1 HI:H_3PO_2 reducing mixture was added to the cooled solution. The flask containing the solution was immediately incorporated into a distillation apparatus with a 25-ml volumetric flask as receiver. The receiving flask,

which contained 15 ml of 30 per cent NH_4OH, was cooled in an ice bath. The assembled apparatus was flushed with nitrogen gas at 200 ml/min for 5 min before the solution was heated and then for 20 min while the flask was heated on a hot-plate with surface temperature of 185°C. (It was found necessary to maintain close control of the hot-plate temperature. In one case when the temperature decreased to 165°C, none of the sulfur was recovered.) The receiving flask was removed, 1 ml of saturated lead citrate solution added, and the solution diluted to 25 ml. The optical density at 370 mμ was then read immediately in a 5-cm cell.

RESULTS

The sulfur contents of eight samples of high-purity indium from five commercial suppliers are listed in Table 1. The sulfur content was less than 0·05 p.p.m. (by weight), the detection limit, in only one sample. The table also shows the sulfur concentration which each indium sample would contribute to InAs prepared from it, assuming complete retention of sulfur during synthesis. A sulfur content of 1 p.p.m. (by weight) in indium corresponds to a sulfur concentration of $6{\cdot}4\times10^{16}$ cm^{-3} in InAs.

Workers at Consolidated Mining and Smelting Company of Canada have recently determined the sulfur content in samples of their High Purity and Special Research Grade indium, using a colorimetric method somewhat different from the one described here. They report a value of $0{\cdot}4\pm0{\cdot}2$ p.p.m. (by weight),[9] in reasonable agreement with the results for Consolidated indium given in Table 1.

DISCUSSION

The indium samples analyzed vary widely in sulfur content. The samples from three suppliers would contribute sulfur concentrations to InAs of approximately 10^{16} cm^{-3}, comparable to the total donor concentrations in the purest InAs reported. The other samples, however, would contribute much lower concentrations. In view of this wide variation—which might also characterize different lots of indium from the same supplier—the present results certainly do not prove that indium is now the major source of donor impurities in high purity InAs. However, they do make it seem probable that in many cases a significant proportion of the donors in InAs originates in the indium. Furthermore, these results also indicate that zone-refined samples of InSb with total impurity concentrations of the order of 10^{14} cm^{-3} are currently being prepared from indium which contributes sulfur concentrations of the order of 10^{16} cm^{-3}.

Table 1. *Sulfur content of commercial indium*

Supplier	Grade	Lot No.	Sulfur content (p.p.m. by weight)	Sulfur conc. contributed to InAs (cm^{-3})
United Mineral and Chemical Corp.	99·999 per cent	12-137	0·36	$2{\cdot}3\times10^{16}$
Consolidated Mining and Smelting Co. of Canada	Vac. melted high purity	H	0·25	$1{\cdot}6\times10^{16}$
	Special research	A-57	0·31	$2{\cdot}0\times10^{16}$
	Special research	A-64	0·19	$1{\cdot}2\times10^{16}$
Ohio Semiconductors, Inc.	In-3	—	0·28	$1{\cdot}8\times10^{16}$
	In-3	—	0·11	$0{\cdot}7\times10^{16}$
Indium Corp. of America	99·999 per cent	—	0·05	$0{\cdot}3\times10^{16}$
American Smelting and Refining Co.	Special high purity	CE-12	$<0{\cdot}05$	$<0{\cdot}3\times10^{16}$

Further evidence for indium as a significant source of impurities in InAs is provided by recent data of WHELAN *et al.*[7] on the purity of arsenic prepared by treatment with molten lead. By using radioactive-tracer techniques, these workers found that when arsenic was condensed into and then distilled from lead at 600°C, the sulfur and selenium concentrations were reduced by factors of approximately 10^3 and 10^4, respectively. They were able to prepare arsenic containing sulfur and selenium at concentrations corresponding to 2×10^{13} cm^{-3} and 5×10^{14} cm^{-3}, respectively, in InAs. If arsenic were the major source of impurities in high-purity InAs, use of arsenic purified by the lead treatment should result in a very large reduction in the impurity concentration in InAs. However, HARMAN *et al.*[1] reduced the impurity concentration by factors of only 3 to 4 when they prepared the compound by combining indium ("99·999+ per cent") from Consolidated Mining and Smelting Company first with arsenic purified by sublimation in hydrogen and then with sublimed arsenic further purified by distillation from a lead–arsenic solution at 675°C. The donor concentration in ingots prepared from the sublimed arsenic was approximately 5×10^{16} cm^{-3}, while samples cut from two ingots made with lead-treated arsenic had donor concentrations of $1{\cdot}3\times10^{16}$ cm^{-3} and $1{\cdot}5\times10^{16}$ cm^{-3}, respectively, and Hall mobilities of 63,000 cm^2/V-sec at 80°K. These results strongly suggest that the indium was the source of a major portion of the donor impurities in the two ingots prepared from lead-treated arsenic, particularly in view of the fact that the samples of Consolidated indium analyzed in the present investigation would contribute sulfur concentrations between $1{\cdot}2\times10^{16}$ cm^{-3} and $2{\cdot}0\times10^{16}$ cm^{-3} to InAs prepared from them.

(In interpreting data on the purity of InAs, contamination must not be neglected as a possible source of impurities. For example, silicon originating in quartz boats and tubes may be a significant donor impurity, as ARNOLD *et al.*[10] have shown it to be in GaAs. However, the degree of silicon contamination should be considerably less for InAs, which melts at 942°C, than for GaAs, which melts at 1240°C.)

On the basis of the data discussed, an attempt should be made to obtain InAs of higher purity than now available by preparing the compound from indium which has been shown analytically to contain a sufficiently low sulfur concentration. At present, the analytical method described above is satisfactory for this purpose, since its limit of sensitivity corresponds to a sulfur concentration in InAs of 3×10^{15} cm^{-3}. The present results also suggest that the selenium content of commercial indium be investigated analytically, since selenium —like sulfur—is not effectively removed from InAs by zone refining. Finally, the possibility of obtaining InP of higher purity by preparing the compound from low-sulfur indium should also be investigated.

REFERENCES

1. T. C. HARMAN, E. P. STAMBAUGH and H. L. GOERING, *Semiconductor Symposium of the Electrochemical Society, October* 1956. Unpublished data.
2. R. H. HARADA and A. J. STRAUSS, *J. Appl. Phys.* **30,** 121 (1959).
3. E. SCHILLMANN, *Z. Naturf.* **11**a, 463 (1956).
4. T. C. HARMAN, *J. Electrochem. Soc.* **103,** 128 (1956).
5. J. B. MULLIN, *J. Electronics* **4,** 358 (1958).
6. A. J. STRAUSS, *J. Appl. Phys.* **30,** 559 (1959).
7. J. M. WHELAN, J. D. STRUTHERS and J. A. DITZENBERGER, *Semiconductor Symposium of the Electrochemical Society, October* 1959. Unpublished data.
8. C. L. LUKE, *Analyt. Chem.* **21,** 1369 (1949).
9. G. H. TURNER. Private communication.
10. E. ARNOLD, J. BLACK and S. WEISSBERGER, *Semiconductor Symposium of the Electrochemical Society, October* 1959. Unpublished data.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 135–143. Printed in Great Britain

THE INTERPRETATION OF THE STATIONARY AND TRANSIENT BEHAVIOUR OF REFRIGERATING THERMOCOUPLES

J. E. PARROTT

Research Laboratory, Associated Electrical Industries, Aldermaston Court, Aldermaston, Berkshire

(*Received* 20 *October* 1959; *in revised form* 25 *November* 1959)

Abstract—The theory of the stationary and transient behaviour of refrigerating thermocouples is presented in an improved form, the Thomson effect, the temperature variation of electrical resistance and surface heat transfers all being taken into account. It is shown that considerable errors can arise from the use of inappropriate approximate expressions, both in the stationary state and in the calculation of the time constants of the transient. In addition, the errors arising from the neglect of the temperature variation of thermal conductivity are estimated and shown to be small.

Résumé—On présente une théorie améliorée en prenant en considération l'effet Thomson, la variation avec la température de la résistance électrique et le transfert de chaleur en surface. On montre que des erreurs considérables peuvent provenir de l'utilisation des approximations impropres aussi bien pour l'état stationnaire que dans les calculs des contacts de temps de l'état transitoire. On montre aussi que les erreurs provenant de l'omission de la variation avec la température de la conductivité thermique sont faibles.

Zusammenfassung—Die Theorie des stätionären und nichtstationären Verhaltens von kühlenden Thermoelementen wird in verbesserter Form dargeboten, wobei der Thomson-Effekt, die Temperaturabhängigkeit des elektrischen Widerstandes und der Wärmeübergang durch die Oberfläche berücksichtigt werden. Es wird gezeigt, dass beträchtliche Fehler durch den Gebrauch ungeeigneter Näherungen entstehen können und zwar sowohl im stationären Zustand als auch bei der Berechnung der Zeitkonstanten. Zusätzlich werden die Fehler, die sich durch die Vernachlässigung der Temperaturabhängigkeit der Wärmeleitfähigkeit ergeben, abgeschätzt; es zeigt sich, dass sie klein sind.

1. INTRODUCTION

IT WAS found in the course of attempting to analyse the stationary and transient behaviour of refrigerating thermocouples that the existing theories were inadequate for this purpose. The object of this paper is to describe the more complete theoretical treatment which enables information about the physical properties of the semiconductors comprising the thermocouple to be deduced.

Two principal experimental arrangements were involved. In the first, the reduction in the temperature of the junction is measured as a function of current passed through it in various situations of thermal loading. In the second, a heater is attached at the junction, and the refrigerating effect is balanced by the power dissipated in the heater, the temperature of the junction being kept constant. In the first experiment, the time dependence on switching the current on and off was also measured. These are the conditions to which the theory to be described is supposed to apply.

The first theory of the refrigerating thermocouple was given by ALTENKIRCH[1] and somewhat more refined accounts have been given by GEHLHOFF *et al.*[2], IOFFE[3] and others. In the earlier accounts very considerable simplifications were made. In particular, all external heat transfers except those due to the thermal load at the junction were neglected, as also were the Thomson effect and the variation of electrical and thermal conductivity and Peltier coefficient with temperature. The transient behaviour has been described in this degree of approximation by STILBANS *et al.*[4]

2. NOTATION

T difference in temperature of semiconductor and its surroundings
κ thermal conductivity
τ Thomson coefficient
σ electrical conductivity
α temperature coefficient of resistance
H outer heat conductance
c heat capacity of the semiconductor
ρ density
p perimeter
A cross-sectional area
j current density
Π Peltier coefficient
r_c contact resistance of unit area
h outer conductance of junction per unit area
S Seebeck coefficient
Γ heat capacity of contact per unit area
l length
w figure of merit

3. DIFFERENTIAL EQUATIONS AND BOUNDARY CONDITIONS

The theoretical model under consideration is shown in Fig. 1. For simplicity it has been assumed that the electrical and thermal properties of the p- and n-arms of the thermocouple are identical except for the absolute Peltier and associated thermoelectric coefficients which are assumed equal and opposite. This means that only one half of the model has to be considered in the calculations.

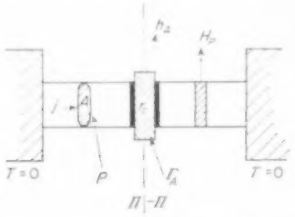


FIG. 1. Refrigerating thermocouple.

The differential equation of heat flow is discussed in CARSLAW and JAEGER[5]. In the calculations to follow, the problem is reduced to one in one dimension only, for which case the appropriate equation is

$$\kappa\frac{\partial^2 T}{\partial x^2}-\tau j\frac{\partial T}{\partial x}+\left(\frac{\alpha j^2}{\sigma}-\frac{Hp}{A}\right)T+\frac{j^2}{\sigma}=c\rho\frac{\partial T}{\partial t} \quad (1)$$

The meanings of the principal symbols used above and later in the paper are given in Section 2.

The first term describes the effects of heat conduction within the material, the second the Thomson effect. The first coefficient of T takes into account the change of electrical conductivity with temperature, whilst the second describes the heat transfers from the surface of the conductor. The last term on the left-hand side is the resistance heating. The right-hand side describes the absorption of heat in warming up the semiconductor. The equation contains no term describing the change of thermal conductivity with temperature as this would produce non-linearity and make an analytical solution impossible. All that can be done here is to employ an average value in the temperature range under consideration. In the Appendix an estimate is made of the errors involved in this approximation.

The boundary conditions are derived from the experimental situation where the outer contacts to the thermocouple are maintained at a constant temperature $T=0$. The situation at the junction will depend on the nature of the experiment. Where heat to balance the refrigeration is supplied, this is also kept at $T=0$, and the calculation is directed to determining the heat input needed to maintain this. If, however, the change in temperature of the junction as a function of current is being measured, then the boundary condition is

$$2\kappa\frac{\partial T}{\partial x}-2\Pi j+r_c j^2-[h+2(S+\tau)j]T=\Gamma\frac{\partial T}{\partial t}. \quad (2)$$

This describes the heat balance at the junction. Thus the first term describes the heat conducted in from the semiconductor, the second is the Peltier effect, and the third is the contact resistance heating; h is the outer conductance of the junction (per unit area of cross-section of the conductors) and so describes the thermal load, whilst $2(S+\tau)j$ (S is the Seebeck coefficient) takes account of the change of Peltier coefficient with temperature $d\Pi/dT=S+\tau$. Γ, on the right-hand side, is the heat capacity (per unit area of conductor) of the contact material.

The initial condition for the time-dependent solutions will, in general, assume that there has been a current j_0 flowing long enough for the thermocouple to come to equilibrium, and that at

$t = 0$, this current is abruptly changed to a new value j. In actual fact, the experiments were restricted to either $j_0 = 0$ or $j = 0$, i.e. switching on from zero current, or off to zero current. Also, the full time-dependent solutions have only been obtained for a simplified differential equation, though the time constants of the transient are presented for the more general case described above.

4. STEADY-STATE SOLUTIONS

The differential equation can be conveniently re-written in the following form

$$\frac{d^2T}{dx^2} - 2b\frac{dT}{dx} + a^2T + d = 0 \qquad (3)$$

where

$$b = \frac{\tau j}{2\kappa},\ a^2 = \frac{\alpha j^2}{\sigma\kappa} - \frac{Hp}{\kappa A},\ \text{and}\ d = \frac{j^2}{\sigma\kappa}.$$

This has the general solution

$$T + \frac{d}{a^2} = A\exp(k_1 x) + B\exp(k_2 x)$$

where $k_1, k_2 = b \pm \sqrt{(b^2 - a^2)}$, the k's being real or complex depending on whether $b^2 - a^2$ is positive or negative.

It is possible to distinguish two conditions, corresponding to (1) producing a fall in temperature, and (2) supplying heat to maintain a steady temperature.

Case 1. This is concerned with the reduction in temperature as a function of current. The boundary condition at $x = 0$ will be re-written as

$$\frac{dT}{dx} - f - gT = 0 \qquad (4)$$

where $f = \dfrac{2\Pi j - r_c j^2}{2\kappa}$, $g = \dfrac{h + 2(S+\tau)j}{2\kappa}$.

The value of T at the junction T_0, is given by

$$T_0 = \frac{\dfrac{d}{a^2}\sqrt{(b^2-a^2)}[e^{-bl}\operatorname{sech} l\sqrt{(b^2-a^2)} - 1] + \left(\dfrac{bd}{a^2} - f\right)\tanh l\sqrt{(b^2-a^2)}}{\sqrt{(b^2-a^2)} - (b-g)\tan l\sqrt{(b^2-a^2)}} \qquad (5a)$$

where $b^2 - a^2 > 0$, or

$$T_0 = \frac{\dfrac{d}{a^2}\sqrt{(a^2-b^2)}[e^{-bl}\sec l\sqrt{(a^2-b^2)} - 1] + \left(\dfrac{bd}{a^2} - f\right)\tan l\sqrt{(a^2-b^2)}}{\sqrt{(a^2-b^2)} - (b-g)\tan l\sqrt{(a^2-b^2)}} \qquad (5b)$$

where $a^2 - b^2 > 0$.

If bl and $l\sqrt{(|a^2-b^2|)}$ are small, it is useful to expand these equations in power series. To the second power of the arguments this gives the result

$$T_0 = \frac{2\Pi^2\sigma}{\kappa}\left\{\frac{y^2\left(1 - \dfrac{\eta y}{3}\right) - y(1-\xi y)\left[1 - \dfrac{\beta}{3} + \dfrac{y^2}{3}\eta^2(\zeta - 1)\right]}{1 + (\nu + \mu y)\left[1 - \dfrac{\beta}{3} + \dfrac{y^2}{3}\eta^2(\zeta - 1)\right]}\right\} \qquad (6)$$

where a new reduced current variable $y = jl/2\Pi\sigma$ has been introduced, and

$$\beta = \frac{Hpl^2}{\kappa A},\ \eta = \frac{\Pi\tau\sigma}{\kappa},\ \xi = \frac{r_0\sigma}{l},\ \zeta = \frac{4\kappa\alpha}{\sigma\tau^2},$$

$$\nu = \frac{hl}{2\kappa},\ \mu = \frac{\Pi(2S+\tau)\sigma}{\kappa}.$$

Further approximating by neglecting terms higher than y^2 in the numerator and y in the denominator leads to

$$T_0 = \frac{2\Pi^2\sigma}{\kappa}\left\{\frac{y^2 - y(1-\xi y)\left(1 - \dfrac{\beta}{3}\right)}{1 + (\nu + \mu y)\left(1 - \dfrac{\beta}{3}\right)}\right\}. \qquad (7)$$

This is the usual expression for T_0 with a correction for surface heat transfer (measured by ν and

β), and contact resistance (measured by ξ). A further discussion of these expressions will appear after the next section.

Case 2. This corresponds to the experimental situation where heat is supplied in such quantities as to maintain a zero temperature difference across the thermocouple, i.e. $T = 0$ at $x = 0$. The heat supplied is given by

$$q_0 = 2\Pi j - r_c j^2 - 2\kappa \frac{\partial T}{\partial x} \tag{8}$$

where $\partial T/\partial x$ is calculated from the solution of (3). The result is

$$q_0 = -2\kappa\left\{\frac{d}{a^3}\sqrt{(b^2-a^2)}\,\frac{[e^{-bl}\operatorname{sech} l\sqrt{(b^2-a^2)}-1]}{\tanh l\sqrt{b^2-a^2)}} + \left(\frac{bd}{a^2}-f\right)\right\} \tag{9a}$$

where $b^2-a^2 > 0$, or

$$q_0 = -2\kappa\left\{\frac{d}{a^2}\sqrt{(a^2-b^2)}\,\frac{[e^{-bl}\sec l\sqrt{(a^2-b^2)}-1]}{\tan l\sqrt{(a^2-b^2)}} + \left(\frac{bd}{a^2}-f\right)\right\} \tag{9b}$$

where $a^2-b^2 > 0$.

As in Case 1 this can be expanded when the arguments bl and $l\sqrt{(a^2-b^2)}$ are small. To the second power

$$q_0 = -\frac{4\Pi^2\sigma}{l}\left(\frac{y^2\left(1-\frac{\eta y}{3}\right)}{1-\frac{\beta}{3}+\frac{y^2\eta^2}{3}(\zeta-1)} - y(1-\xi y)\right) \tag{10}$$

is obtained, where the symbols have the same meaning as in Case 1. Further approximating by neglecting terms above y^2 in the numerator and y in the denominator gives

$$q_0 = -\frac{4\Pi^2\sigma}{l}\left(\frac{y^2}{(1-\beta/3)} - y(1-\xi y)\right). \tag{11}$$

The uses and limitations of these equations will now be discussed in terms of numerical values typical of good thermoelectric materials such as bismuth telluride. Such values are:

$\Pi = 6\times10^{-2}$ V $\quad \tau = 1{\cdot}5\times10^{-4}$ V deg^{-1}
$S = 2\times10^{-4}$ V deg^{-1} $\quad \kappa = 2\times10^{-2}$ W cm^{-1} deg^{-1}
$\alpha = 5\times10^{-3}$ deg^{-1} $\quad \sigma = 10^3$ mho-cm^{-1}

The corresponding values of η, γ and μ are $\eta = 0{\cdot}45$, $\zeta = 17{\cdot}8$ and $\mu = 1{\cdot}65$. The dimensionless figure of merit, $w = \Pi^2\sigma/\kappa T$ (T is the absolute temperature in this equation) is 0·6.

Suppose the length of the thermocouple is made $l = 1$ cm, then the current for $y = 1$, i.e. $2\Pi\sigma/l$ is 120 A/cm^2. Further, a contact resistance r_c of $2{\cdot}8\times10^{-4}$ Ω-cm^2 will be assumed so that $\xi = 0{\cdot}28$.

The values of the outer heat conductance H will depend on the ambient conditions. In free air, H might be approximately 2×10^{-3} W/cm^2 per deg, and in vacuum 2×10^{-4} W/cm^2 per deg or less. Thus if the thermocouple rods are 0·4 cm edge, β is unity in air and $< 0{\cdot}1$ in vacuum. It will be seen that in the former case its effect may be very considerable. Reasonable values of ν will be 10^{-1} in air and 10^{-2} in vacuum.

The expected values of bl and $l\sqrt{(a^2-b^2)}$ will now be discussed. A rough calculation using (8) shows that the maximum cooling would occur at $y = \frac{1}{2}$ if the contact resistance were zero and there were no surface heat transfers. In the experimental situations envisaged, y will be kept less than unity since most of the interesting phenomena can be seen within this range.

Equations (6) and (10) will lead to incorrect values of T_0 and q_0 if bl and $l\sqrt{(a^2-b^2)}$ are not small. Taking bl first, it is seen that $bl = \eta y = 0{\cdot}45\,y$, and it is clear that for $y \sim 1$, the use of (6) or (10) would give an error of several per cent. The situation is more serious still for $l\sqrt{(a^2-b^2)}$. By substitution,

$$l\sqrt{(a^2-b^2)} = \sqrt{[y^2\eta^2(\zeta-1)-\beta]}$$

and neglecting β, $= 1{\cdot}85\,y$.
In this case, serious errors would arise for all values of y greater than about 0·25.

The consequences of the Thomson effect and the variation of electrical resistivity with temperature are illustrated in the figures. Fig. 2 shows the

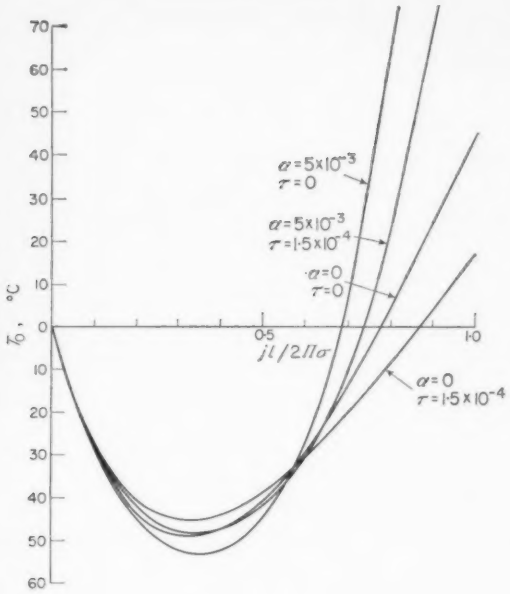


FIG. 2. Cooling as a function of current: $\Pi = 6\times10^{-2}$ V; $\sigma = 10^3$ mho/cm; $S = 2\times10^{-4}$ V deg^{-1}; $\kappa = 2\times10^{-2}$ W cm^{-1} deg^{-1}; $r_c\sigma/l = 2{\cdot}8\times10^{-1}$; $\Pi^2\sigma/\kappa T = 0{\cdot}6$.

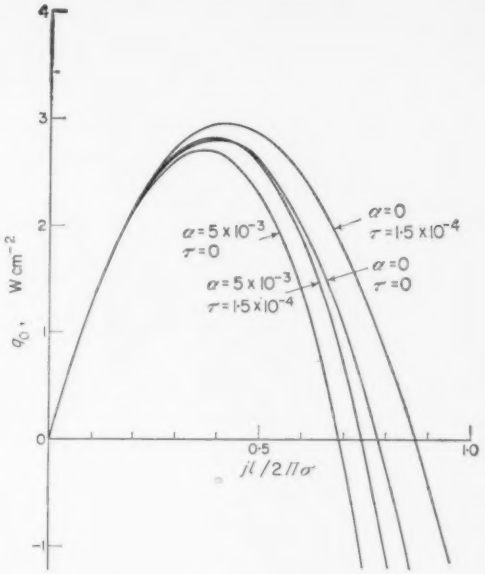


FIG. 3. Refrigerating capacity as a function of current: $\Pi = 6\times10^{-2}$ V; $\sigma = 10^3$ mho/cm; $S = 2\times10^{-4}$ V deg^{-1}; $\kappa = 2\times10^{-2}$ W cm^{-1} deg^{-1}; $r_c\sigma/l = 2{\cdot}8\times10^{-1}$; $\Pi^2\sigma/\kappa T = 0{\cdot}6$.

cooling as a function of current. The simplest formula assumes $\alpha = 0$, $\tau = 0$. Here a noticeably asymmetric curve is obtained, due to the variation of the Peltier coefficient with temperature. Taking account of τ but neglecting α increases the asymmetry since instead of putting $d\Pi/dt = S$ the correct result $d\Pi/dt = (S+\tau)$ is now used. This results in a slight diminution of the cooling at its maximum value also, though this is partly compensated by the bulk Thomson effect which is helpful. Taking account of α, but neglecting τ leads to a substantially greater cooling at small currents but less at large currents. This is because when the thermocouple is almost completely below $T = 0$, the resistance heating is reduced by α, but, as shown, as the larger part of the thermocouple is above $T = 0$, the resistance heating increases. When both α and τ are considered, the curve tends to that for $\alpha = 0$ when the current is small and to that for $\tau = 0$ when the current is large. Over a wide range, the values are not very different from those obtained by neglecting both.

Fig. 3 shows the heat extraction as a function of current. The curve for $\alpha = 0$, $\tau = 0$ is now a parabola. Taking account of the Thomson effect increases the heat extraction since the temperature gradient and current are in such a sense that the Thomson effect refrigerates near the junction. The temperature coefficient of resistance has the reverse effect, as might be expected, and again the curve taking account of both τ and α is close to that where they are both neglected, tending to curve for $\tau = 0$ at high currents.

The relation of the approximate formulae (6), (7), (10) and (11) to the correct ones is shown in Figs. 4 and 5. These show that Approximation I, using equations (6) and (10), leads to large errors at high currents for the values of τ and α chosen; for smaller values they would be more valuable. Approximation II, using equations (7) and (10), is better, but this is due partly to the fact that the effects of τ and α cancel out.

To summarize the results of this section, it may be said that for normal values of τ and α, the effects of the correction are large only for currents rather greater than those for which either T_0 or q_0 is a maximum. Furthermore, the second-order

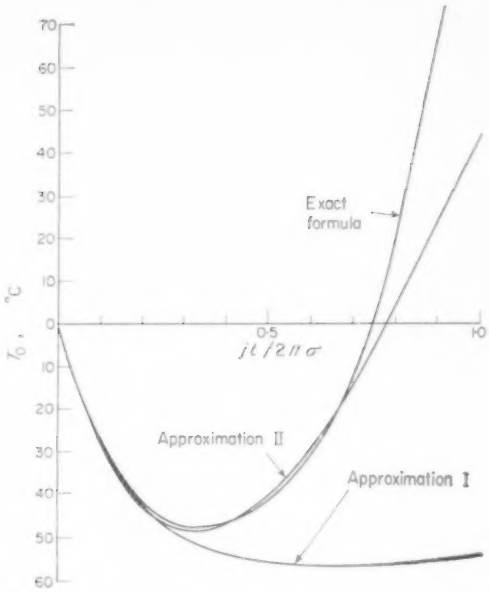


FIG. 4. Use of approximate formulae for cooling: $\Pi = 6\times10^{-2}$ V; $\sigma = 10^3$ mho/cm; $S = 2\times10^{-4}$ V deg^{-1}; $\kappa = 2\times10^{-2}$ W cm^{-1} deg^{-1}; $r_c\sigma/l = 2{\cdot}8\times10^{-1}$; $\Pi^2\sigma/\kappa T = 0{\cdot}6$.

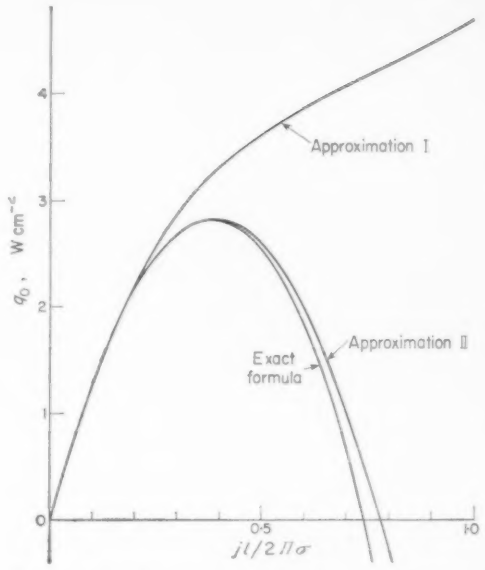


FIG. 5. Use of approximate formulae for refrigerating capacity: $\Pi = 6\times10^{-2}$ V; $\sigma = 10^3$ mho/cm; $S = 2\times10^{-4}$ V deg^{-1}; $\kappa = 2\times10^{-2}$ W cm^{-1} deg^{-1}; $r_c\sigma/l = 2{\cdot}78\times10^{-1}$; $\Pi^2\sigma/\kappa T = 0{\cdot}6$.

approximations give worse agreement with the exact formulae, for the reasons given in the previous paragraph.

5. TRANSIENT BEHAVIOUR

On changing the current passing through a refrigerating thermocouple, some time elapses before a new equilibrium is attained. It would be valuable to be able to use the curve of this transient to determine some of the physical properties of the semiconductor, and for this reason a time-dependent solution of (1) was sought. This can be done either by the use of Laplace transforms or as an eigenvalue problem. The first is the most suitable for obtaining a complete solution, but is a lengthy process, whilst the second is useful as a much quicker method of determining the time constants of the transient.

The eigenvalue approach shows that the solution has the general form

$$T_0 + \frac{d}{dt^2} = \sum_{i=0}^{\infty} A_i \exp(-t/\tau_i) \qquad (12)$$

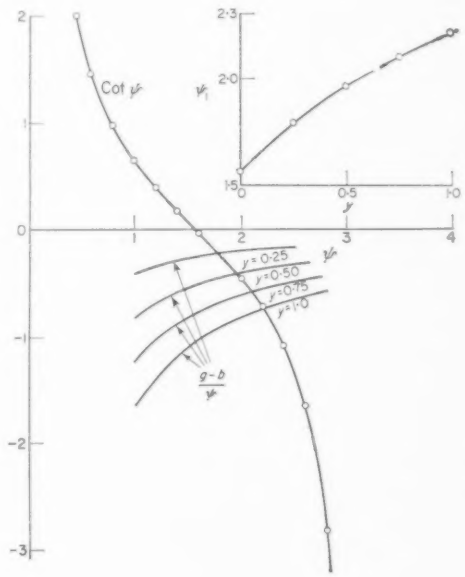


FIG. 6. Determination of ψ_1.

where
$$\frac{1}{\tau_i}=\frac{\kappa}{c\rho l^2}[\psi_i^2-(a^2-b^2)l^2] \qquad (13)$$

and the ψ_i's are the roots of the transcendental equation

$$\beta\psi-\frac{1}{\psi}[\beta l^2(a^2-b^2)-b+g]=\cot\psi \qquad (14)$$

$$\beta=\frac{\Gamma}{2c\rho l}.$$

Fig. 6 shows how the roots of (14) are affected by changes in the current. Here a simplification has been made by neglecting β, which is the same as assuming that the heat capacity of the junction is zero. It is seen that ψ_1 increases steadily with increasing current. Then, inserting the appropriate value of ψ_1 into (13), using the same expressions for a^2 and b^2 as were used in Section 4, and putting $c\rho/\kappa = 70$ cm^2/sec^{-1}, the variation of τ_1 with current shown in Fig. 7 is found. There is an initial fall in τ_1 followed by a rise at higher currents. The fall is due to changes in ψ_1, the rise to the increase of $(a^2-b^2)l^2$. The equation predicts an infinite value of τ_1 at a certain value of the current. Presumably this corresponds physically to a situation in which the temperature rises without limit owing to the rise in dissipation at constant current when α is positive. The situation is, in this approximation, unstable above certain values of current.

The effect of a large value of β is to raise the time constant at small currents as extra time is needed to extract the heat from the contact material in addition to the material of the thermocouple. However, there will be a more rapid change of ψ with current owing to the $\beta(a^2-b^2)l^2$ term. All of these effects have been shown by experiment.

The coefficients A_i are still unknown, however,

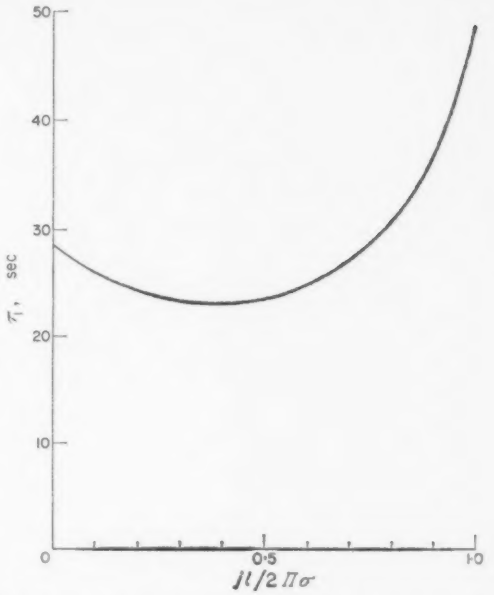


FIG. 7. Variation of τ_1 with current: $1/\tau_1 = \kappa/c\rho l^2[\psi_i^2-(a^2-b^2)l^2]$; $c\rho/\kappa = 70$ cm^2 sec^{-1}; $a^2 = \alpha j^2/\sigma\kappa$; $b^2 = \tau^2 j^2/4\kappa^2$.

and the solution of a simplified equation was obtained by R. BULLOUGH using Laplace transforms. The simplified differential equation is

$$\kappa\frac{\partial^2 T}{\partial x^2}+\frac{j^2}{\sigma}=c\rho\frac{\partial T}{\partial t} \qquad (15)$$

whilst the boundary and initial conditions are those described earlier. The general solution is

$$T_0(t)=\frac{l^2d/2-fl}{1+gl}+\frac{2l}{(1+gl)}\sum_{n=1}^{\infty}\frac{\exp\left(\frac{-\kappa\psi_i^2 t}{c\rho l^2}\right)}{\psi_i^2}\times$$

$$\times\frac{\left[f_0(1+\beta\psi_i^2)-f(1+gl)+\frac{d_0 l}{2}(gl-\beta\psi_i^2)\right]\psi_i\sin\psi_i+\left(\frac{d_0 l}{2}-f_0\right)\psi_i^2\cos\psi_i+(d-d_0)l(1+gl)(1-\cos\psi_i)}{(1+gl-\beta\psi_i^2)\cos\psi_i-\psi_i(2\beta+1)\sin\psi_i} \qquad (16)$$

where the subscript zero means that the value of $j = j_0$ appropriate to $t < 0$ must be used.

For the case of switching on, the simpler expression

$$T_0(t)=\frac{I^2\lambda l/2-fI}{1+gl}+2I\sum\frac{\exp\left(\frac{-\kappa\psi_i^2t}{c\rho l^2}\right)}{\psi_i^2}\frac{dl(1-\cos\psi_i)-f\psi_i\sin\psi_i}{(1+gl-\beta\psi_i^2)\cos\psi_i-\psi_i(2\beta+1)\sin\psi_i} \tag{17}$$

is obtained, which is the same as that of STILBANS,[4] whilst equation (17) for the switching-off transient

$$T_0(t)=\frac{2I}{(1+gl)}\sum\frac{\exp\left(\frac{-\kappa I\psi_i^2t}{c\rho l^2}\right)}{\psi_i^2}\times$$
$$\times\frac{[f_0(1+\beta\psi_i^2)+\frac{d_0l}{2}(gl-\beta\psi_i^2)]\psi_i\sin\psi_i+\left(\frac{d_0l}{2}-f_0\right)\psi_i^2\cos\psi_i-d_0l(Hgl)(1-\cos\psi_i)}{(1+gl-\beta\psi_i^2)\cos\psi_i-\psi_i(2\beta+1)\sin\psi_i} \tag{18}$$

The equation determining the ψ_i's is also simpler than (12), being

$$\beta\psi-\frac{gl}{\psi}=\cot\psi. \tag{19}$$

It is found, however, that the full expressions (13) and (14) are necessary to explain the experimental time constants.

6. CONCLUSIONS

The calculations presented provide a more realistic basis for comparison of theory with experiment in the study of refrigerating thermocouples. The temperature dependence of the thermal conductivity is not taken into account, but the considerations given in the Appendix suggest that the errors arising from this are not serious. It is also possible that the variation of contact resistance with temperature might be of importance in some cases. If this is so, it is a simple matter to insert a term which will take this into account.

The main difficulty in using the formulae arises from the number of parameters involved, and the fact that the approximate formulae will often be inaccurate. However, by combining steady-state and transient observations with a very few subsidiary measurements, it should be possible to obtain a self-consistent picture. An analysis of experimental results using this theory will shortly be presented.

Acknowledgements—The author would like to thank Dr. R. BULLOUGH for carrying out the calculations described in Section 5, and Dr. T. E. ALLIBONE, F.R.S., Director of the Laboratory, for permission to publish this paper.

APPENDIX

Variation of thermal conductivity with temperature

If the effect of the temperature variation of coefficient of thermal conductivity β is included, equation (1) becomes

$$\kappa\frac{\partial^2T}{\partial x^2}+\kappa\beta\left(\frac{\partial T}{\partial x}\right)^2+\tau j\frac{\partial T}{\partial x}+\left(\frac{\varkappa j^2}{\sigma}-\frac{Hp}{A}\right)T+$$
$$+\frac{j^2}{\sigma}=c\rho\frac{\partial T}{\partial t}. \tag{A1}$$

This equation is non-linear and not soluble analytically. However, the simpler equation

$$\kappa\frac{\partial^2T}{\partial x^2}+\kappa\beta\left(\frac{\partial T}{\partial x}\right)^2+\frac{j^2}{\sigma}=0 \tag{A2}$$

can be solved, and can be used to estimate the effects of neglecting β in the full equation. The general solution of (A2) is

$$\kappa T+\kappa\frac{\beta T^2}{2}=\frac{j^2}{\sigma}\left(\frac{x^2}{2}+Ax+B\right) \tag{A3}$$

i.e.

$$T=-\frac{1}{\beta}\left\{1\pm\sqrt{\left[1-\frac{2j^2\beta}{\sigma\kappa}\left(\frac{x^2}{2}+Ax+B\right)\right]}\right\}.$$

The minus sign before the square root is the one to be used, as can be seen by expanding the root. If the first

two terms are taken one finds

$$T = -\frac{j^2}{\sigma\kappa}\left(\frac{x^2}{2} + Ax + B\right) \quad \frac{j^4\beta}{2\sigma^2\kappa^2} \times \times \left(\frac{x^2}{2} + Ax + B\right)^2. \tag{A4}$$

The first term is that obtained when β is neglected and the second term is the first-order correction.

The values of A and B depend on the boundary condition. To give a parallel with the real problem, put $T = 0$ when $x = l$, and $-\kappa(\partial T/\partial x) = W$ when $x = l$. The value of T at $x = l$ is then approximately

$$T_l = -\frac{j^2 l^2}{2\sigma\kappa}\left(\frac{2W\sigma}{l_j{}^2} - 1\right)\left[1 + \frac{j^2\beta l^2}{4\sigma\kappa}\left(\frac{2W\sigma}{l_j{}^2} - 1\right)\right] \tag{A5}$$

Now, approximating a little more yields

$$T_l \approx -\frac{j^2 l^2}{2\sigma\kappa}\left(\frac{2W\sigma}{l_j{}^2} - 1\right)\left(1 - \frac{\beta T}{2}\right)$$

$$\approx -\frac{j^2 l^2\left(\frac{2W\sigma}{l_j{}^2} - 1\right)}{2\sigma\kappa\left(1 + \frac{\beta T}{2}\right)} \tag{A6}$$

This shows that an accuracy to $0(\beta^2 T^2/4)$ can be obtained by using the average value of thermal conductivity in expressions in which it appears. The lattice thermal conductivity is approximately inversely proportional to the absolute temperature, so, at 300°K, $\beta \approx 33 \times 10^{-3}$. Since T will not generally exceed 50°C, the error involved is less than 1 per cent.

REFERENCES

1. E. Altenkirch, *Phys. Z.* **12,** 920 (1911).
2. P. O. Gehlhoff, E. Justi and M. Kohler, *Abh. braunschw. wiss. Ges.* **2,** 149 (1950).
3. A. F. Ioffe, *Semiconducting Thermoelements and Thermoelectric Cooling*. Infosearch, London (1957).
4. L. S. Stilbans and N. A. Fedorovich, *Zh. tech. Phys., Moscow* **28,** 489 (1958).
5. H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*. Clarendon Press, Oxford (1947).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 144–147. Printed in Great Britain

RADIATION LIMITED PbS CELLS

R. L. WILLIAMS*

Canadian Armament Research and Development Establishment

(*Received* 14 *December* 1959)

Abstract—The noise of Eastman Kodak PbS chemically deposited cells has been measured in an excess-radiation environment. Under this condition, and that of no excess radiation, the same equation describes the noise of the cells.

Résumé—Le bruit de fond des cellules de PbS d'Eastman Kodak, déposées chimiquement, étaient mesurées dans un rayonnement abondant. Dans ces conditions aussi bien que lorsqu'il n'y a pas de rayonnement environnement, le bruit de fond est décrit par la même équation.

Zusammenfassung—Das Rauschen von chemisch niedergeschlagenen Eastman-Kodak-PbS-Zellen wurde in einer Umgebung mit Strahlungsüberschuss gemessen. Sowohl unter dieser Bedingung als auch ohne überschüssige Strahlung beschreibt dieselbe Gleichung das Rauschen der Zellen.

1. INTRODUCTION

THE phrase "radiation noise" has been used to describe the minimum noise output of cooled photoconductive detectors. The noise is predicted on the basis that a cell can follow fluctuations in the background radiation which are slower than the response time of the cell. FELLGET,[1] MOSS[2] and CLARK JONES[3] have considered this problem. In each case the fluctuations of the photon beam were considered and the cell response to such fluctuations taken as the noise of a detector working in a non-thermal equilibrium environment of excess radiation. PETRITZ[4] points out that for a photoconductor there are fluctuations in the current-carrier generation and recombination rates, and derives an equation for the noise on this basis. As VAN VLIET[5] points out, the same theory can be applied for the excitation of current carriers by photons or phonons. The expression for the noise in each case is the same if the equation is written down in terms of parameters of the semiconductor. The mean-square voltage fluctuation appearing across a photoconductor carrying a current and having a series resistance R_L is given by

$$V^2 = 4I^2 \frac{R_c R_L^2}{R_c + R_L} \frac{\tau \Delta f}{N(1+w^2\tau^2)} \qquad (1)$$

In this expression τ is the lifetime of the current carriers, N is the total number of current carriers produced by photons and phonons, W is the angular frequency and Δf the bandwidth. The expression derived by FELLGET, MOSS and CLARK JONES differs from the above expression, having a factor of 2 instead of 4. This is equivalent to neglecting the recombination of the photon-excited carriers.

The expression for g–r (generation–recombination) noise, as the above type is called, has been shown by LUMMIS and PETRITZ[6] to describe quite adequately the g–r noise of PbS chemically deposited photoconductive cells. The work was performed at room temperature without excess radiation.

The expression for the photoconductive voltage, for the conditions of operation mentioned above, is given by PETRITZ[7] to be

$$Vpc = I \frac{R_c R_L}{R_c + R_L} \frac{\phi_s \tau \gamma}{N} \frac{1}{(1+W^2\tau^2)^{1/2}} \qquad (2)$$

The applicability of this equation to PbS cells for an r.m.s. sinusoidal photon signal ϕ_s/sec has been established by WOODS[8]. The factor γ includes the quantum efficiency and any losses at the surface of the cell.

2. EXPERIMENTAL PROCEDURE

The low-frequency ratio of equation (1) to (2) is

* Now with RCA Victor Company Ltd., Research Laboratories, Montreal, Quebec, Canada.

$$\frac{Vn^2}{Vpc} = 4\frac{IR_cR_L}{(R_c+R_L)}\frac{\Delta f}{\gamma\phi_s} \qquad (3)$$

Writing $A = \gamma\phi_s/\Delta f$,

$$A = \frac{Vpc}{Vn^2}\frac{4IR_cR_L}{(R_c+R_L)}$$

In the experiments to be described ϕ_s and Δf were kept constant and values of A were evaluated using equation (3) by substituting the measured values of the parameter for various conditions of excess radiation. The theories of PETRITZ and VAN VLIET suggest that A should be the same regardless of the amplitude of the background radiation intensity.

The experiments were performed at the temperature of solid CO_2. At this temperature by controlling the wavelength of the light, g–r noise can be produced under conditions of excess radiation.[9] Room-temperature operation was attempted first, as the same wavelength noise spectrum dependence found by WILLIAMS[9] at CO_2 temperatures was thought to exist at the higher temperature. Using a germanium filter and a 1000 W lamp whose filament was just visibly red, a change from the g–r to the $1/f$ spectrum was found as soon as cell resistances had changed by a factor of approximately two. This result, combined with the earlier work by WILLIAMS suggests that wavelengths in the 1·0 to 2·8-μ region produce the g–r to $1/f$ noise change at room temperature.

The test arrangement used to make the noise measurements is depicted in Fig. 1. The cell chamber was made of brass, each end having a vacuum "O" ring seal. Various excess radiation levels were obtained by controlling the filament voltage of the 12 W lamp. As only low voltages were applied to the lamp, adequate long-wavelength intensities could be obtained without filtering the radiation. The photoconductive signal was produced by applying a variable frequency, 75 V square wave to the neon bulb. The amplitude of the light was held small by covering the lamp with a number of sheets of paper. The cell resistance change produced by the neon light was never greater than 1 per cent. The neon light was enclosed in a small wire-mesh cage to reduce stray electromagnetic radiation.

The vacuum connection was made with a vacuum valve mounted at the experimental

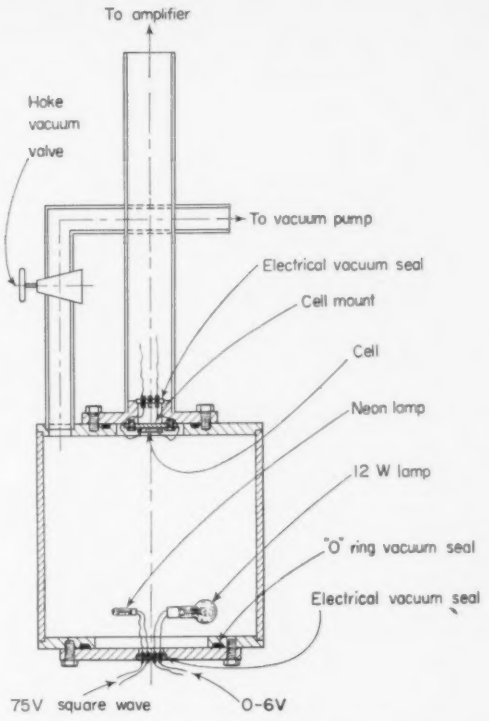


FIG. 1. Experimental chamber used in the investigation of cell noise with background radiation.

chamber. With the valve closed and the chamber and valve cooled the cell chamber was completely isolated and the no-background condition could be ideally realized. The system proved to be not sufficiently vacuum tight to permit this, but the vacuum valve and associated vacuum tubing proved to be a very effective radiation trap. With the vacuum valve and chamber cooled, closing and opening the valve produced no observable change in resistance. This corresponded to a $\Delta R/R$ of less than 1 per cent.

The noise measurements were recorded using a system similar to that described previously.[9] The photoconductive frequency response was measured with the noise recording system for the purpose of curve fitting. Both Vn^2 and Vpc^2 were plotted as a function of frequency and the two sets of points compared to obtain the best agreement in relative amplitude. Then the theoretical curve $(1+W^2\tau^2)^{-1}$ was drawn through the points. This is illustrated in

Fig. 2. As can be seen, good agreement is obtained between the theoretical curve and the experimental noise and photoconductive points. The ately apparent that the A values for these cells do not change with background radiation. The separate symbols indicate points taken in a single cool-

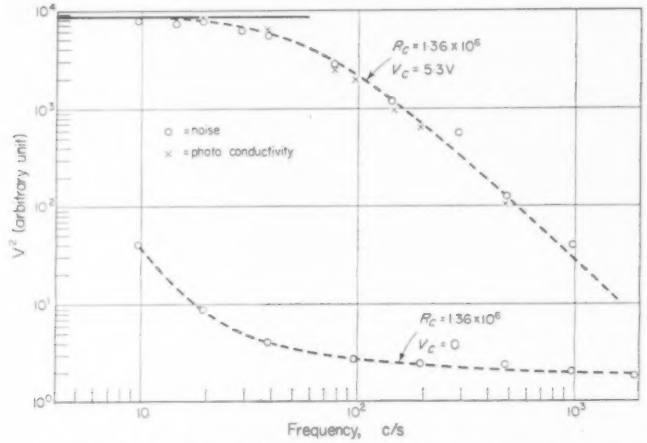


FIG. 2. Comparison of the noise spectrum and frequency response of a PbS cell. The zero current noise illustrates its relative magnitude to the g–r noise.

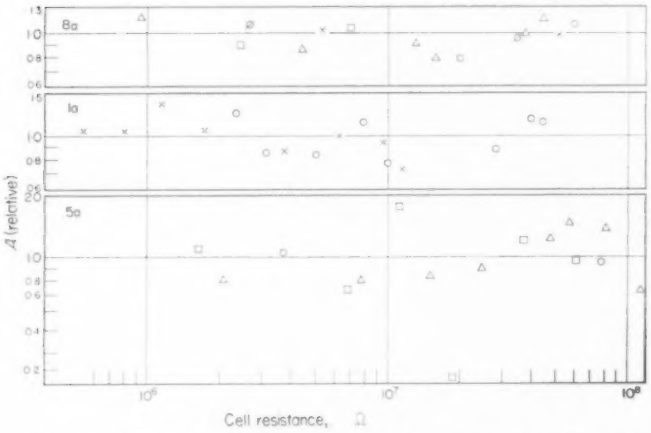


FIG. 3. Comparison of the parameter A for various background-radiation levels—cells 8a, 1a and 5a.

value of Vpc to be used in the evaluation of A was taken as the square root of the low-frequency saturation value of V^2pc.

3. EXPERIMENTAL RESULTS

The results of the three most extensive sets of measurements are given in Fig. 3. It is immediing of the system. About 1 hr was required for all the measurements taken for each point.

The relative magnitude of the g–r and $1/f$ noise components was found to be quite consistent in a particular lot of cells ordered. The cells used for the readings of Fig. 3 were all from one lot, and no cell tested from this group had $1/f$ noise with no

background radiation at CO_2 temperatures. Of all other cells checked, about fifteen in number, only three were useable, as the rest had excess $1/f$ noise under all conditions of radiation. For these three the $1/f$ noise was small only for one or two days and then it became large. The reason for this changing magnitude of $1/f$ noise was not determined.

No results of the present experiments at CO_2 temperatures are in conflict with the earlier results by WILLIAMS[9]. A change to dominance by $1/f$ noise was found for high incandescent-lamp voltages. This is expected as higher filament voltages result in not only larger numbers of photons but a shift of the wavelength distribution to shorter wavelengths. The range of cell resistances over which A values could be evaluated was limited by this change from g–r to $1/f$ noise.

4. DISCUSSION

The results of the experiments reported indicate that for photons in the range 1–2·7μ, photon and phonon excitation are statistically indistinguishable. They also support the theories of PETRITZ and VAN VLIET which include all the processes taking place.

As the noise of a photoconductive cell depends on the random recombination of carriers as well as fluctuation in the generation rate caused by the background radiation, the term "radiation noise" is not a complete description of the noise output of the cell. The phrase "radiation limited" is suggested to describe the condition of the cell. BURNSTEIN and PICUS[10] describe this condition by saying that the cell is a "background limited" infrared photoconductor or that the cell is in the "blip" condition of operation.

The calculation of photon backgrounds to estimate the difference between a cell sensitivity and the theoretical limit is unnecessary. A simple comparison of the measured cell sensitivity with the theoretical value based on equations (1) and (2) serves equally well.

For a photoconductor which has its resistance controlled by radiation, the number of current carriers is related to the number of background photons incident per second, ϕ_b, by the expression $N = \phi_b \tau \gamma$. Using this expression in equations (1) and (2) the signal-to-noise ratio of a g–r radiation-noise-limited device is

$$\frac{S}{N} = \frac{Vs}{Vn} = \gamma^{1/2}\phi_s(4\phi_z\Delta f)^{-1/2}$$

It is immediately apparent that a reduction of the background radiation improves the sensitivity of a photoconductor under these conditions.

5. REFERENCES

1. P. FELLGET, *J. Opt. Soc. Amer.* **39,** 1970 (1949).
2. T. S. MOSS, *J. Opt. Soc. Amer.* **40,** 693 (1950).
3. R. CLARK JONES, *Advances in Electronics* Vol. V, Chap. 1. Academic Press, New York (1953).
4. R. L. PETRITZ, *A Photoconductivity Conference* Chap. 2. Wiley, New York (1956).
5. K. M. VAN VLIET, *Proc. I.R.E.* **46,** 1004 (1958).
6. F. L. LUMMIS and R. L. PETRITZ, *Phys. Rev.* **105,** 502 (1957).
7. R. L. PETRITZ, *Phys. Rev.* **104,** 1508 (1956).
8. J. F. WOODS *Phys. Rev.* **106,** 235 (1957).
9. R. L. WILLIAMS, *Canad. J. Phys.* **37,** 841 (1959).
10. E. BURNSTEIN and G. S. PICUS, Paper presented at the Infrared Information Symposium, February 1958; R. L. PETRITZ, *Proc. I.R.E.* **37,** 1458 (1959)

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 148–156. Printed in Great Britain

MINORITY-CARRIER EFFECTS IN CHEMICALLY DEPOSITED PbS PHOTOCONDUCTIVE FILMS

R. L. WILLIAMS*

Canadian Armament Research and Development Establishment

(*Received* 14 *December* 1959)

Abstract—Using the PEM and Hall effect it has been established that the minority-carrier diffusion length in chemically deposited PbS cells is at least as large as the film thickness. The relative magnitude of the PEM voltage when cells are illuminated from the front and back surfaces indicates that the recombination process between 300°K and 200°K is governed by surface phenomena. The cell time constant, however, is determined by trapping processes.

Résumé—On montre par des mesures d'effet P.M.E. et d'effet Hall que la longueur de diffusion des porteurs minoritaires dans les cellules de PbS déposées chimiquement est au moins aussi grande que l'épaisseur des couches. La valeur relative de la tension P.M.E. lorsque les cellules sont illuminées sur les surfaces de face ou de l'envers indique que les processus de recombinaison à 300°K et à 200°K sont régis par des phénomènes de surface. La constante de temps est cependant déterminée par le processus de piègeage.

Zusammenfassung—Unter Benutzung des PEM- und des Hall-Effektes wurde festgestellt, dass die Diffusionslänge der Minoritätsträger in chemisch niedergeschlagenen PbS-Zellen mindestens so gross ist wie die Dicke des Films. Das Verhältnis der PEM-Spannungen im Falle einer Bestrahlung der Zellen von vorne bzw. von der Rückseite zeigt, dass der Rekombinationsprozess swischen 300°K und 200°K von Oberflächen-Phänomenen abhängt. Die Zeitkonstante der Zellen wird jedoch von Trap-Prozessen bestimmt.

INTRODUCTION

A GREAT deal of work has been undertaken to determine the factors controlling the sensitivity of PbS films. For many years much effort was used in attempting to establish the role, if any, of barriers. Work by GIBSON,[1] MAHLMAN[2] and SLATER[3] indicated that barrier modulation could be responsible for the photoconductive response. However, theory by PETRITZ[4] and the experimental work of WOODS[5] gave strong evidence that the PC (photoconductivity) effect was primarily a carrier-density phenomena.

In all models it was necessary to have a recombination process and only recent papers have devoted any effort to this problem. The work of KLASSEN and BLOK[6] suggested the need of two levels to account for the recombination process. A single recombination level was suggested by PETRITZ *et al.*[7] but one which was different from the level controlling the resistance. In the latter paper, it was suggested that in the sensitization process the minority-carrier lifetime became very small through the introduction of sites into which the carriers were readily trapped. These centers had, it was suggested, small cross sections of recombination for holes, and thus one obtained long lifetimes.

PHOTOELECTROMAGNETIC EFFECT (PEM)

GARTNER[8] has developed expressions for the optically produced change in conductance and the PEM short-circuit current and open-circuit voltage. The restriction of a very large absorption coefficient has been relaxed in GARTNER's development. The PEM open-circuit voltage, V_s, for a crystal with a diffusion length much greater than the sample thickness and with surfaces controlling

* Now with RCA Victor Company Ltd. Research Laboratories, Montreal, Quebec, Canada.

the combination, is given by

$$V_s = \frac{q\theta Il}{\sigma K} \times$$

$$\times \left[\frac{KS_2(1+S_1/S_2e^{-\kappa})+S_1S_2(1-e^{-\kappa})}{S_1+S_2+S_1S_2} + 1 - e^{-\kappa} \right] \tag{1}$$

In this expression q is electronic charge, θ the Hall angle, σ the conductivity of the sample, I the number of photons incident per sec per cm^2, K the product of the absorption coefficient κ and the sample thickness d, $S_1 = s_1 d/D$ and $S_2 = s_2 d/D$; in which s_1 and s_2 are, respectively, the surface recombination velocities of the front and back surface of a sample whose ambipolar diffusion coefficient is D.

This expression reduces to a simple form if K is much greater than one, and recombination at the back surface dominates, $S_2 \gg S_1$, viz.

$$V_s = \frac{q\theta Il}{\sigma}\left[\frac{1}{1+S_1}\right] \tag{2}$$

If the surface recombination velocity of the front surface is greatest, $S_1 > S_2$ and

$$V_s = \frac{-q\theta Il}{\sigma}\left[\frac{S_2}{S_1(1+S_2)} - \frac{1}{K(1+S_2)}\right] \tag{3}$$

It is seen that for dominant front-surface recombination the sign of the PEM voltage depends on the relative magnitudes of the two terms of expression (3). If the second term is dominant then the ratio of the open-circuit PEM voltage when the photons are incident on the front surface, V_{sf}, to that obtained by illuminating the same sample from the back, under identical conditions, V_{sb}, (S_1 becomes S_2 and S_2 becomes S_1) is

$$\frac{V_{sf}}{V_{sb}} = -\frac{1}{K} = -\frac{1}{kd} \tag{4}$$

It is seen that the ratio of the PEM voltages gives immediately a value of the absorption coefficient if the sample thickness is known, or it could give relative values for different wavelengths if d is unknown.

The effect of trapping on diffusion has been considered by JONSCHER[9] and his conclusions are that, if steady-state or low frequencies are considered, diffusion distances are very little altered. GARTNER's theory should then be applicable to direct-current PEM measurements, even in the presence of trapping.

EXPERIMENTAL PROCEDURE

Hall measurements

The applicability of the theory of the Hall effect to PbS Ektron detectors has been considered by LUMMIS and PETRITZ[10] and WOODS.[5] Their model of the film structure is accepted and the expressions they quote are used to calculate mobilities and the product pd, the hole-carrier density times the thickness of the film. An Applied Physics Corporation vibrating-reed electrometer was used to measure Hall voltages. Its input impedance, being greater than $10^{15}\ \Omega$, was sufficiently high to insure effective open-circuit measurements for all cell resistances encountered. The instrument sensitivity of 1 mV full scale was more than adequate for the measurement of the Hall voltages encountered. The resistance of the film was measured directly, the contacts being assumed to give no errors in this measurement.

The PEM effect for polycrystalline films

For the PEM measurement the cell equivalent circuit was taken to be composed of two parts; a current generator described by PEM theory in parallel with the resistance the cell would have if barriers were absent, and the junction resistance in series with this combination. As the junction resistance is dominant, it is the value obtained when the cell resistance R_c is measured. With this equivalent circuit the PEM voltage V_{pem} measured by an instrument having an input resistance R_m would be $V_{\text{pem}}\, R_m/(R_m+R_c)$. This was verified experimentally. In general, open-circuit voltages were measured so that the effect of the junction resistance was nullified.

The physical picture envisaged is that each crystallite can be treated as a separate PEM generator and the cell PEM voltage is the sum of a large number of such generators. As the barrier resistance is large compared to that of the crystallites, the PEM voltage is not laterally shorted out.

The PEM voltages were measured with a Hewlett-Packard 425A direct-current microvolt-meter. The instrument, used with its input resistance of 1 MΩ removed, had a measured input resistance of 350 MΩ. For cells with high resistances the

vibrating-reed electrometer was used to measure the PEM voltages. A gain of 100 in sensitivity was obtained by applying the full meter-output voltage across a 25 mV full-scale Minneapolis-Honeywell recorder with suitable attenuating resistances. Used in this fashion 20 or 30 μV could be readily detected. For measurements of both PEM and Hall voltages the signals were recorded with a Minneapolis-Honeywell instrument.

The cells were shaped by placing a mask, having the desired configuration, over the cell and sandblasting away the unwanted cell areas. The procedure found most satisfactory for making electrical contact to the side arms was as follows. Gold was evaporated across approximately one half of the side arm, the plastic coating first being removed. The gold area was sufficiently large that a fine wire could be conveniently glued to the area with silver paint. These contacts proved to have photovoltages not appreciably greater than the proper cell contacts and showed little rectifying characteristics. As Hall voltages were measured under conditions of high illumination large photovoltages made measurements difficult.

Most of the cells used were from a batch of 10×10 mm units which were shaped to give a central portion 5×10 mm with 1 mm side arms. Two 4×4 mm cells were used but approximately the same proportions were maintained.

Hall and PEM effects were performed using four wavelength regions: Bausch and Lomb filters were employed for 1·25 and 2·1 μ with a band pass of approximately one-fifth the center wavelength. A sodium lamp with a heat filter to isolate the lines at 0·5 μ was used in limited cases; its intensity being rather small restricted the range of usefulness. As a general short-wavelength source, limited to wavelengths less than 0·8 μ, a tungsten lamp and a heat filter were employed. In all cases no focusing was used, to insure as uniform an illumination as possible.

For absolute magnitude measurements, the light output of a Westinghouse 75 W projector lamp and the appropriate filter was compared to a calibrated bolometer.

The Hall measurements under illumination were taken using a small mirror mounted just before the cryostat window to deflect the light beam 90°, i.e. the light beam was incident between the magnet poles.

The cryostat used for the measurements consisted of a liquid-air chamber from which the sample mount was suspended by copper–nickel tubing. The thermal conductance of the tubing was such that temperatures near that of liquid air could be reached and a heater wound on the sample mount could readily raise the temperature to 300°K. A thermocouple was employed to measure temperatures. The light beam used in the experiments was incident through a glass window of the vacuum jacket.

EXPERIMENTAL RESULTS

(i) *General*

As PbS photoconductive films are polycrystalline, a check was made to see how well the PEM voltage obeyed theory. The effect was linear with magnetic field between 0·8 and 9·6 kG, the latter field being used in practically all the measurements. The saturation of the PEM voltage with H, as predicted by KURNICK and ZITTER,[11] of $\mu B(1+ +\mu^2B^2)^{-1/2}$ is not expected with the 10 kG field employed and a carrier mobility of 5–600 cm^2/V-sec. The PEM voltage was found to be linear with light flux for the range of photon signals from $1{\cdot}1 \times 10^{14}$ to $1{\cdot}6 \times 10^{15}$ photons per sec. As the results to be quoted range over light-flux intensities of the order of 50 the generally consistent pattern of results obtained is good evidence of the general validity of the theory. The interpretation of parameters calculated and the effect of the film structure will be discussed further subsequently.

(ii) *The nature of the PEM effect*

One of the best indications that surfaces are controlling the recombination is through the use of formulae (2) and (3) and by illuminating the cell from the front and back. If one surface recombination is dominant the ratio of the front to the back open-circuit PEM voltages is kd and the signal changes sign on rotation. The PEM voltage was measured, then the cells rotated 180°, and the voltage measured again. The light intensity and magnetic field were not changed in the two experiments. A polarity reversal was found on rotation for all cells and all wavelengths tested: 0·5, 0·8, 1·25 and 2·25 μ. Table 1 lists the results of such measurements. The loss of transmission due to the presence of the glass mounting for one side and the plastic film for the other was considered equal. The

Table 1. *Evaluation of the absorption coefficient by illuminating a cell from the front and the back* (T = 295°K)

λ (μ)	$\phi \times 10^{-14}$ (photons/sec)	kd	$k \times 10^4$ cm^{-1} ($d = 6 \times 10^{-5}$ cm)	Rc (M)
1·25	1·3	1·4	2·7	2·50
1·25	2·8	1·3	2·2	2·45
1·25	6·3	1·5	2·5	2·33
1·25	14	1·5	2·5	2·14
Average $k = 2{\cdot}5 \times 10^4$, SCANLON $k = 3{\cdot}0 \times 10^4$ ($\lambda = 1{\cdot}25\mu$)				2·45
0·5	—	3·8	6·7	2·51
0·5	—	6·0	10·0	
Average $k = 8{\cdot}3 \times 10^4$, SCANLON $k = 5{\cdot}0 \times 10^4$ ($\lambda = 0{\cdot}5\mu$)				

agreement between the average of the calculated values of the K's and those of SCANLON[12] is quite good.

The validity of GARTNER's theory has been checked by BRAND *et al.*[13] but as details of their results have not been published a simple experiment was performed on a sample of Ge to check the validity of equations (3) and (4). The sample thickness was reduced to $9{\cdot}6 \times 10^{-3}$ cm or about 1/10th of the sample's measured diffusion length. With one surface etched and the other ground, the experiment mentioned above for PbS was performed to determine k. A value of $5{\cdot}2 \times 10^3$ cm^{-1} was obtained which is 26 per cent lower than the value determined by DASH and NEWMAN[14]. As a rather broad wavelength region was used a smaller value of k is expected from the PEM measurement. This result and the fact that the signal reversed sign when the sample was rotated 180° are taken as good confirmation of the theoretical expressions.

MEASUREMENT OF L_x

Most of the PEM measurements were performed by illuminating the cell from the front surface and calculating an "effective diffusion length" L_x defined by the equation:

$$V_s = \frac{2\phi B L_x \times 10^{-8}}{pdW}$$

The factor L_x/d is equal to the bracketed terms of equations (1–3). In this expression ϕ is the number of photons incident on the cell per second, W is the width of the cell and pd the product of the hole density times the thickness of the film. The factor 2 arises as the electron and hole mobilities have been assumed equal as suggested by the work of ALLGAIER and SCANLON[15]. The product pd was evaluated from Hall measurements made under identical conditions of illumination to those of the PEM measurement. The most extensive set of measurements of L_x was performed with cell No. 3 which was a 10×10 mm cell. The results are listed in Table 2. It is immediately apparent that L_x is essentially independent of temperature and light intensity.

Table 2. *Diffusion length L_x for various light-flux levels and temperatures* (*sample* PbS *No.* 3; $\lambda = 1{\cdot}25\ \mu$)

T (°K)	$\phi \times 10^{-14}$ (photons/sec)	$L_x \times 10^6$ (cm)
187	8·5	4·2
192	8·5	6·0
193	8·5	4·0
193	14·5	5·5
193	5·25	7·5
193	4·0	4·0
225	11·5	2·2
225	7·0	1·9
225	3·2	4·0
225	1·9	5·5
297	7·75	2·5
297	3·50	4·0
297	2·15	4·2
297	1·0	4·5
297	0·35	4·2
298	2·2	6·0
314	13	3·2
314	8·0	4·2
314	3·5	4·5

A simple check of the wavelength-dependence of L_x was carried out using cell No. 2. Table 3 lists comparisons of L_x for 2·1 and 1·25 μ at room temperature and approximately 195°K. The values

Table 3. Wavelength variation of the diffusion length L_x (sample No. 2; $H = 9{\cdot}8$ kg)

T (°K)	$\phi \times 10^{-14}$ (photons/sec)	$L_x \times 10^5$ (cm)	λ (μ)
293	11	2·6	2·1
293	2·0	2·3	2·1
295	9·5	0·67	1·25
295	1·25	0·76	1·25
196	11	2·91	2·1
196	2·0	2·8	2·1
194	3·5	0·50	1·25
194	1·1	0·50	1·25
195	4·7	0·55	1·25
195	3·0	0·55	1·25

for k given by SCANLON[12] are: $\lambda = 1{\cdot}25\ \mu$, $k = 3{\cdot}0 \times 10^4$ cm^{-1}; $\lambda = 2{\cdot}1\ \mu$, $k = 1{\cdot}4 \times 10^4$ cm^{-1}. Although the change of Lx with wavelength is not large, it is to be noted that it is always larger for the longer wavelength. L_x being proportional to the bracketed terms of equations (1–3), the behaviour is consistent with equation (3) if the second term in the bracket is dominant. Some variation of L_x with temperature is apparent for this cell but it is not large.

As has been generally found, cells of a particular order have similar characteristics. This was found for cells 1–6 which were the 10 × 10 mm cells. Two 4 × 4 mm cells, 7 and 8, from a different purchase lot, were measured to see if the results were comparable to the 10 × 10 mm cells. Results of measurements on cell 7 can be summarized as follows:

(1) At room temperature the average value of Lx for 1·25 μ radiation was $2{\cdot}7 \times 10^{-5}$ cm (two readings).

(2) The corresponding value at 194°K was $9{\cdot}1 \times 10^{-5}$ cm (nine readings).

(3) The ratio of L_x for 2·1 μ to that at 1·25 μ was 1·3 (293°K).

(4) A sign reversal in the PEM voltage was found at room temperature when the cell was rotated 180°.

(5) The mobility was essentially independent of light flux and wavelength. (This information to be used later.)

HALL MEASUREMENTS

To evaluate L_x of the PEM effect it was necessary to determine pd. This was obtained from Hall measurements. In the course of these measurements it was noted that the mobility was a function of light level. Assuming that a unique relationship exists between the carrier density and the carrier mobility, some crude limits might be put on the value of the diffusion length by measuring the carrier mobility when the carriers are excited by different wavelengths of light. If the diffusion length of the minority carriers is much less than the reciprocal of the absorption coefficient, the optically excited carriers remain essentially where created, i.e. in a layer about k^{-1} cm thick. By contrast, if the diffusion length L_D is long compared to k^{-1} and the thickness of the film, the position of excitation by the use of different wavelengths of light would have little or no effect on the density distribution, the carriers quickly diffusing across the film. Diffusion times will be discussed in a following section.

If the carriers are confined to a layer about k^{-1} cm thick the value pd determined from Hall measurements would be in fact pk^{-1}. As the mobility was known to vary with p, a result to be presented, the pd values obtained for each wavelength of light would have to be multiplied by a factor $(dk)^{-1}$ to compare mobilities for the same carrier density. Or conversely, if μ is plotted against pd, and L_D is much shorter than k^{-1}, the mobility curves would not be expected to coincide, as p would not be simply proportional to the calculated value of pd but in error by different factors proportional to k^{-1}. This assumes, as mentioned above, that μ and p are uniquely related.

The results of the wavelength-mobility measurements are given in Fig. 1. It is immediately apparent that the mobility is a function of p. There is no appreciable difference in the value of μ for a given value of pd indicating that the diffusion length is not less than k^{-1}.

For cell No. 7, as previously mentioned, no variation of mobility with the density of optically excited carrier was found.

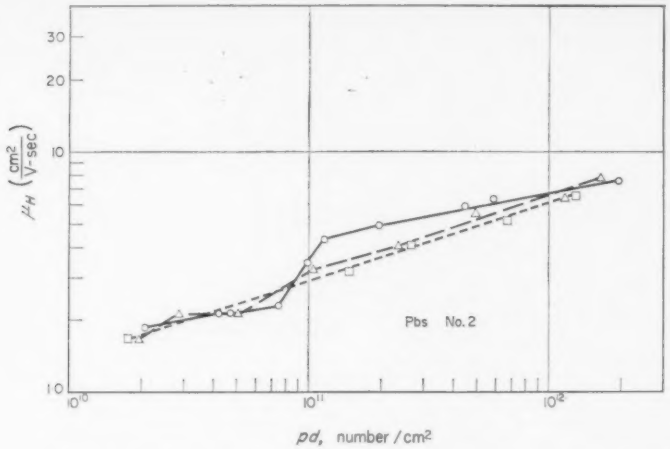


FIG. 1. Mobility of optically excited carriers for three wavelength regions.

○ $T = 195°K$, $\lambda = 2{\cdot}12\mu$
□ $T = 192°K$, $\lambda < 0{\cdot}8\mu$
△ $T = 195°K$, $\lambda = 1{\cdot}25\mu$

The experimental work of WOODS[5] dealt with the change of carrier mobility with light injection. His work established that in the range $\Delta R/R \leqslant 1$ the photo-effects were primarily carrier-density effects. From the results just quoted the mobility changes are quite secondary for $\Delta R/R$ values up to approximately 100.

RECOMBINATION PROCESS—RESPONSE TIME

The results of the PEM measurements indicate that recombination is taking place at the surfaces of the cell. The films are of the order of $0{\cdot}6\,\mu$ thick[16,17] and the time constant should then be controlled by recombination at the surfaces. For a semiconductor which has identical minority- and majority-carrier lifetimes SHOCKLEY[18] and MCKELVEY and LONGINI[19] have developed expressions for the response time. For a thin layer with surface recombination dominant the lifetime becomes $d/2s$. The measured cell time constant is of the order of 3×10^4 so that for a film 6×10^{-5} in. thick s is 0·1 cm/sec. This is a low value of s.

A second case is possible, namely that s is large but that diffusion is limiting the rate of recombination. The response time is then $4d^2/\pi D$. The diffusion coefficient D can be estimated from ALLAGAIER and SCANLON's[15] value for the single-crystal electron mobility. As diffusion is across the film, barriers should not appreciably affect the value of D and it is calculated to be 14 cm^2/sec at room temperature. The cell response time would then be 3×10^{-10} sec, a value 10^{-6} times smaller than observed.

The above points indicate that processes other than simple recombination at the surface are taking place. Further evidence has been provided for this by the work of WOODS[5]. From his analysis of Hall and conductivity data on PbS films it is apparent that minority-carrier lifetimes must be small compared to that of the majority carriers. This leads to two possible schemes. The minority carriers are trapped temporarily in the bulk of the crystallites and, after being thermally ejected from the centers, retrapped, probably many times, the carriers pairs eventually recombining at the surface.

As an alternative to this process, minority carriers could be trapped at the surface and remain there until the eventual recombination with free holes, the majority carriers. If the latter process is the correct one, minority-carrier effects would be essentially complete when the electrons are trapped at the surface. By contrast to this behaviour, if trapping takes place in bulk with eventual recom-

bination at the surface, minority-carrier effects would exist for a time comparable to the majority-carrier lifetime.

HAYNES and HORNBECK[20] derive the following expression for the response time in the presence of minority-carrier trapping:

$$\tau = \tau_g + \tau_r + \tau_g\tau_r/\tau_t$$

where τ_g is the average time spent in traps, τ_r is the lifetime (time in the appropriate band) and may be the sum of a number of smaller intervals, and τ_t is the time before trapping. The expression used for τ_t is related to the number of trapping centers per cm^3, N, the thermal velocity v, and the capture cross section Q, by the expression $\tau_t = (NQv)^{-1}$. The expression for the time constant is then

$$\tau = \tau_g + \tau_r + \tau_g\tau_r NQv$$

For silicon, HAYNES and HORNBECK's results suggested that for moderate light pulses all the traps are saturated when the light is on, i.e. N approaches zero and the time constant is for the first part of the decay $\tau = \tau_g + \tau_r$. As time progresses some of the traps are emptied and N, the number available for trapping, increases until the time constant becomes $\tau_g\tau_r NQv$.

If N is larger than normal carrier concentrations encountered, and the trapping process is dominant, a light pulse will decay with a time constant $\tau_r\tau_g NQv$ and be unaffected by the light-pulse amplitude. The time constant of the device is then determined by minority-carriers effects.

FREQUENCY RESPONSE

To aid in determining the physical picture of the time history of the excited carriers, the frequency response of cell No. 2 was measured for the photoconductive, photovoltaic and photoelectromagnetic effects. The results of these measurements are shown in Fig. 2. The amplitudes of the PEM and PC signals have been arbitrarily made equal for easy comparison of the frequency response. The P–V curve was placed at a convenient amplitude for comparison with the other two curves. The drop in the photoconductive response curve near 3·5 kc/s indicates a cell time constant of 46 μsec.

The photovoltaic response shows no time constant effects out to 12 kc/s, but by contrast the

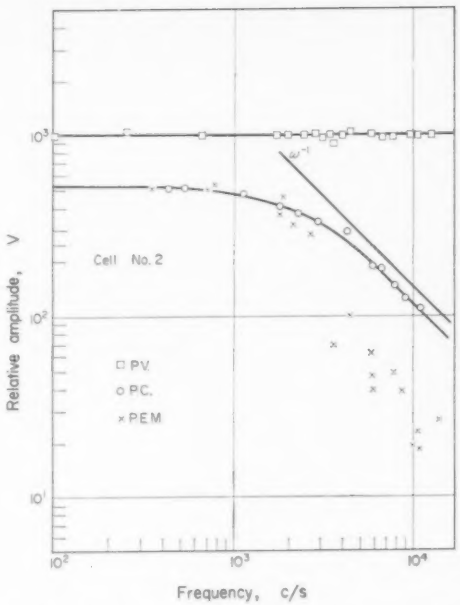


FIG. 2. Frequency response for the photoconductive, photovoltaic and photoelectromagnetic effects.

PEM signal shows an even stronger frequency response than the photoconductivity. The significant feature of these curves is that the PEM has a time-constant effect similar to that of the photoconductivity. The measurement of the PEM signal is difficult; even at low frequencies the magnitude of the PV and PEM signals are comparable, while at 10 kc/s the PEM voltage represents only a few per cent modulation of the photovoltage. The scatter in points near 10 kc/s is a consequence of this.

Light intensities which changed the cell resistance by a factor of nearly 3 were used to measure the response curves of Fig. 2. This is not considered to reduce the validity of the results for the following reasons. Trapping effects such as described by HAYNES and HORNBECK[20] tend if anything to be removed if high light levels are employed. PEM experiments performed with comparable light intensities gave no indication of any changed behaviour.

DISCUSSION

(a) *General*

The concept of trapping as outlined by HAYNES and HORNBECK,[20] coupled with surface domin-

ance of recombination, explains quite satisfactorily the results of the present experiments. The sensitivity of the cells arises from the presence of traps in which the minority carriers spend most of their time, and this delays the eventual recombination at the surface. During the periods that the minority carriers are trapped the majority carriers remain free to provide photoconduction.

The significance of the lack of any evidence for a photovoltaic time constant comparable to the PC value is hard to evaluate. The photovoltaic voltage is the sum of many separate photovoltages at each of the inter-crystalline junctions. For crystallites whose dimensions are of the order of 10^{-4} cm there are approximately 10^8 junctions in a 10×10 mm cell. The problem of amplitude and time constant can be very complex.

The junction mobility has been shown to be modified in a secondary manner by optically induced carrier-density changes.

(b) *Polycrystalline films*

The application of single-crystal theory to polycrystalline films is not without question of validity. The argument presented of each crystallite being a small PEM generator is felt to be valid but correction factors need to be applied to the magnitude of Vs obtained from such crystallites. The agreement between the value of k determined from the ratio of front to back PEM voltages and that determined by SCANLON[12] is felt to be partly fortuitous, but, taking the ratio of front to back PEM signals, common correction factors cancel out.

On the whole, sufficient agreement with theory is found to justify the conclusions drawn but the calculated values of L_x should be treated as order-of-magnitude values. The present experiments indicate that most of the recombination is taking place at the front surface of the cells. This, combined with the sign reversal found when the cell was rotated 180°, suggests that L_x should be equal to k^{-1}, the second term of the bracket in equation (3). For the 10×10 mm cells the value of L_x is about one order of magnitude smaller than k^{-1}. That L_x is smaller is not unexpected, as any correction for the polycrystalline nature of the film would tend to increase the calculated value of L_x. For the 4×4 mm cells L_x was comparable to k^{-1} but this result is not necessarily to be taken as better agreement with theory.

(c) *Low-frequency noise*

The present experiments were undertaken in an effort to find some substance for the arguments used by the author[21] to explain the origin of $1/f$ noise in cells irradiated with short-wavelength radiation. The principle of short diffusion lengths rather than the concept of trapping was used to explain the results. The present experiments suggest that regardless of the position of excitation the surfaces control the recombination. The lack of $1/f$ noise with long-wavelength background radiation cannot then be based on the explanation that recombination is in the bulk when long-wavelength radiation excites carriers there.

ROBERTS[22] has suggested, and WILLIAMS[21] noted, that the energy difference of the different wavelengths might be the significant feature. ROBERTS points to the presence of the peak in sensitivity near 1μ which he observed in PbSe and which also has been found in PbS.[1] He suggests that excitation of this level is the source of $1/f$ noise. Two things are unsatisfactory with the explanation: first the $1/f$ noise could, in certain cases, be produced with radiation of wavelengths longer than $1{\cdot}8\,\mu$[23] and secondly that Eastman Kodak chemically deposited PbS cells do not in general show this $1\,\mu$ peak.[17,24]

Although the trapping effects believed to be present are somewhat akin to the small-diffusion-length arguments used previously,[21] the results reported, unfortunately, do little to clarify the earlier experimental results on the productions of $1/f$ noise by radiation. An explanation is more likely to be found if the source and mechanism of the $1/f$ noise can be determined.

Acknowledgements—The author wishes to acknowledge many helpful discussions with G. GIROUX, and the assistance of M. SHEREBRIN who performed many of the measurements.

REFERENCES

1. A. F. GIBSON, *Proc. Phys. Soc. Lond.* B**64,** 595 (1951).
2. G. W. MAHLMAN, *Phys. Rev.* **103,** 1619 (1956).
3. J. C. SLATER, *Phys. Rev.* **103,** 1631 (1956).
4. R. L. PETRITZ, *Phys. Rev.* **104,** 1508 (1956).
5. J. F. WOODS, *Phys. Rev.* **106,** 235 (1957).
6. F. M. KLASSEN and J. BLOK, *Physica 's Grav.* **24,** 975 (1958).

7. R. L. Petritz, F. L. Lummis, H. E. Sorrows and J. F. Woods, *Semiconductor Surface Physics* p. 229. University of Pennsylvania Press (1957).
8. W. Gartner, *Phys. Rev.* **105,** 823 (1957).
9. A. K. Jonscher, *Proc. Phys. Soc. Lond.* B**70,** 230 (1957).
10. F. L. Lummis and R. L. Petritz, *Phys. Rev.* **105,** 502 (1959).
11. S. W. Kurnick and R. N. Zitter, *J. Appl. Phys.* **27,** 278 (1956).
12. W. W. Scanlon, *J. Phys. Chem. Solids* **8,** 423 (1959); *Phys. Rev.* **109,** 47 (1958).
13. F. A. Brand, A. N. Baker and H. Mette, *Bull. Amer. Phys. Soc.* **2,** 171 (1957).
14. W. C. Dash and R. Newman, *Phys. Rev.* **99,** 1151 (1951).
15. R. S. Allgaier and W. W. Scanlon, *Phys. Rev.* **111,** 1029 (1958).
16. H. E. Spencer, *J. Appl. Phys.* **30,** 927 (1959).
17. H. E. Spencer. Private communication (1959).
18. W. Shockley, *Electrons and Holes in Semiconductors* p. 323. Van Nostrand, New York (1950).
19. J. P. McKelvey and R. L. Longini, *J. Appl. Phys.* **28,** 634 (1957).
20. J. R. Haynes and J. A. Hornbeck, *Photoconductivity Conference* p. 321. Wiley, New York (1956).
21. R. L. Williams, *Canad. J. Phys.* **37,** 841 (1959).
22. D. H. Roberts, *J. Electron. Contr.* **5,** 256 (1958).
23. R. L. Williams, *Solid-State Electron.* **1,** 144 (1960).
24. A. Smith and D. Dutton, *J. Opt. Soc. Amer.* **48,** 1007 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 157–158. Printed in Great Britain

NOTE

A masking technique for vacuum evaporation

(*Received* 9 *December* 1959)

WITH increasing emphasis on miniaturization and high-frequency performance of semiconductor devices, vacuum-evaporation techniques for the deposition of small electrodes have come into prominence. One example of a sophisticated application of this technique is, for instance, the mesa high-frequency transistor,[1] developed by the Bell Telephone Company.

Experimentally, one of the major problems of this technique is the development of proper evaporation masks through which the electrodes are to be deposited; this may be an expensive and time-consuming process, especially if a variety of geometries are to be experimented on.

In the following, a laboratory technique is described which makes it possible essentially to replace intricate masks by a simple structure during evaporation and a subsequent etching procedure.

Fig. 1(a) shows the principle of the mask. The shield, in the form of a wedge, prevents the vapor from source *A* from reaching side *b* of the slice and similarly that from source *B* from reaching side *a*. By making the distance *x* small, the requirement for very close contact between shield and slice becomes less critical, as is obvious from the geometry of the arrangement. In practice the shield is supported at its edges so that it can move freely in a vertical plane. It will thus adjust itself properly even if the substrate is slightly slanted.

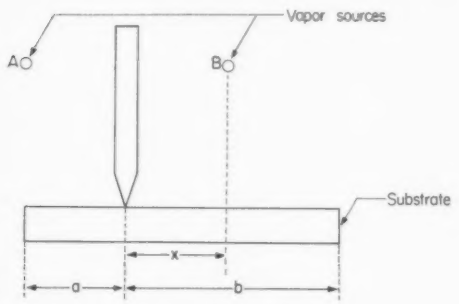


FIG. 1(a)

Fig. 1(b) shows a slice with two evaporated layers separated by 0·4 mil, obtained by the above technique. The distance between the layers is determined by the width of the wedge in contac with the slice.

In order to define the desired electrode configuration after alloying in the evaporated layers and at the same time to make a pedestal, a mask of wax is applied covering the strip separating the alloys, as shown in Fig. 1(c). Finally the slice is diced up along the broken lines. Alternatively, the slicing is performed first and the individual units are masked and etched to produce the desired geometry. A photograph of one particular configuration obtained in this way is shown in Fig. 2.

The simple mask shown in Fig. 1(a) can be refined to allow simultaneous evaporation to produce a number of parallel layers on one slice, using several wedge-shaped shields spaced suitably apart as shown schematically in Fig. 3(a), for the case of three sets of evaporated layers. In this case the height of the shields and their distance apart must be properly related to the height of the vapor sources above the substrate. As indicated

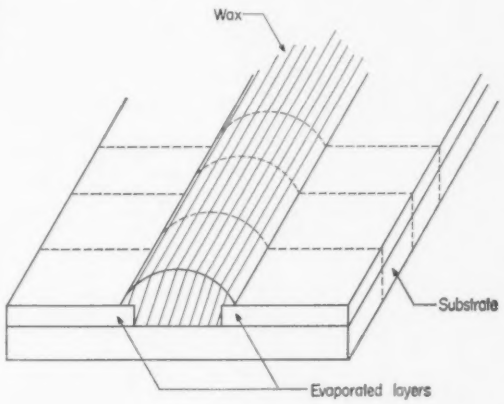


FIG. 1(c)

in the figure, there will be regions of overlapping layers depending in extent on the geometry used. However, this is of no significance since usually the required width of layers *a* and *b* is so small that most of the region between two wedges, including the overlapping layers, will be etched away in any case. The multiple shield arrangement requires, in addition to the wedge-shaped shields, a shield *D* to prevent mutual contamination of the vapor sources.

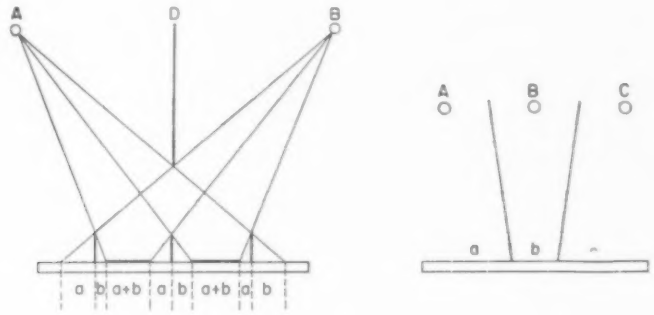


FIG. 3.

A simple extension of the multiple shield arrangement consists of using deposition from three vapor sources for fabrication of three-electrode devices. Fig. 3(b) illustrates the principle which encompasses delineation of the middle electrode by the wedges and that of the outer electrodes by etching. As before, the width of the wedges will determine the spacing between the electrodes.

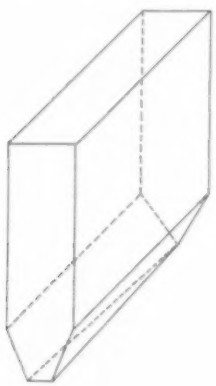


FIG. 4.

Another wedge shape found to be useful for development work is shown in Fig. 4. This wedge produces a continuously varying separation between evaporated layers on one slice, making it possible to study the effect of electrode separation under otherwise equal conditions.

The above methods obviously do not exhaust the possibilities of the technique described and can, with some ingenuity, be applied to obtain a variety of other electrode configurations.

REFERENCE

1. C. A. LEE, *Bell. Syst. Tech. J.* **35**, 23 (1956).

W. RINDNER

Research Division
Raytheon Company
Waltham, Mass.

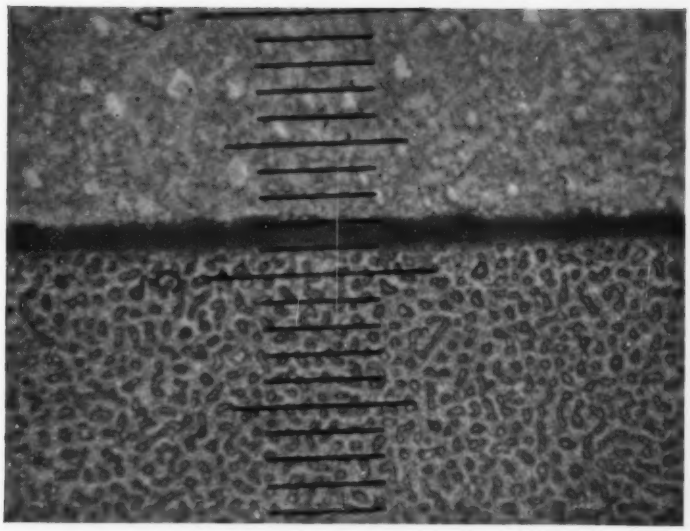


FIG. 1(b). 0·4 mil/scale division.

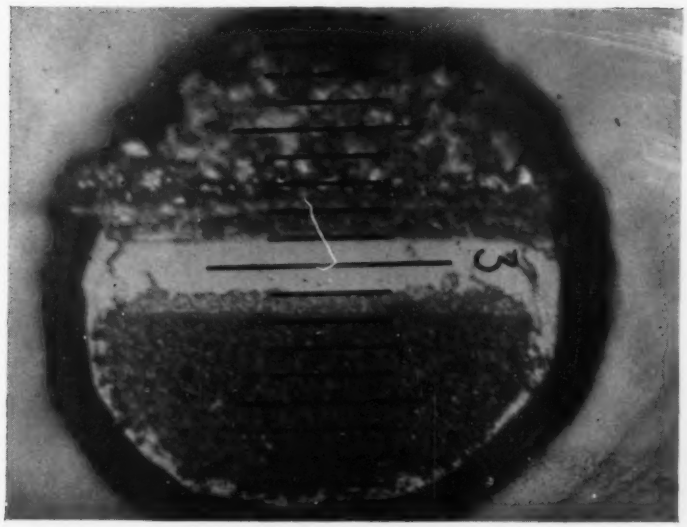


FIG. 2. 0·4 mil/scale division.

[*facing p.* 158

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 159–163. Printed in Great Britain

ELECTRICAL CONDUCTIVITY OF HIGH VANADIUM PHOSPHATE GLASS

MOTOSUKE MUNAKATA

Electrotechnical Laboratory, Nagata-cho, Chiyoda-ku, Tokyo

(*Received* 28 *November* 1959)

Abstract—The electrical conductivity of high vanadium phosphate glass is studied as a function of its composition. The conductivity varies with the concentration of vanadium 4+, if the concentration of total vanadium is kept constant, and has a maximum for a certain critical value of V^{4+}/V_{total}. From these results, it is concluded that the conduction is due to valence–electron exchange between V^{4+} and V^{5+}, similar to the case in Fe_3O_4.

When the third component of the glass, namely basic metal oxide, is varied, a simple relation between the maximum conductivity of the glass and the "cationic field strength" of the metal is found. By this means, it has proved possible to obtain semiconducting glass with a room-temperature resistivity of the order of 10^3 Ω-cm.

Résumé—La conductivité électrique est étudiée en fonction de la composition du verre. La conductivité varie avec la concentration du vanadium 4+ si la concentration totale du V est maintenue constante et elle a un maximum pour une certaine valeur critique de V^{4+}/V_{total}. On conclut de ces résultats que la conduction est due à l'échange d'électrons de valence entre V^{4+} et V^{5+}, analogue au cas du Fe_3O_4.

Lorsqu'on fait varier la troisième composante du verre, notemment des oxydes alcalins, on obtient une relation simple entre la conductivité maximum du verre et la "force du champ du cation" du métal. On prouve ainsi qu'il est possible d'obtenir des verres semiconducteurs dont la résistivité à température ordinaire est de l'ordre de 10^3 Ω-cm.

Zusammenfassung—Die elektrische Leitfähigkeit von hochgradigem Vanadiumphosphat-Glas wird als Funktion der Zusammensetzung studiert. Die Leitfähigkeit ändert sich mit der Konzentration von Vanadium 4+, und, wenn die Gesamtkonzentration des Vanadiums konstant gehalten wird, und hat ein Maximum für einen bestimmten kritischen Wert von V^{4+}/V_{total}. Aus diesen Ergebnissen wird geschlossen, dass die Leitfähigkeit durch den Valenzelektronen-Austausch zwischen V^{4+} und V^{5+} verursacht wird, ähnlich wie im Falle des Fe_3O_4.

Wenn die dritte Komponente des Glases, nämlich das basische Metalloxyd, geänder wird, so lässt sich eine einfache Beziehung zwischen der maximalen Leitfähigkeit des Glases und der "Feldstärke der Kationen" des Metalls finden. Hierdurch hat es sich als möglich erwiesen, halbleitendes Glas mit einem Widerstand in der Grössenordnung von 10^3 Ωcm bei Zimmertemperatur zu erhalten.

1. INTRODUCTION

THERE have been several reports on the semiconducting properties of glasses containing vanadium oxide as the principal constituent.[1–3] Among them, BAYNTON *et al.*[3] report that the system V_2O_5–P_2O_5–BaO (sometimes Na_2O added) shows *n*-type electrical conductivity of 10^{-4}–10^{-7} Ω^{-1} cm^{-1}.

They also report that, for the same system of components, the conductivity is determined by V_2O_5 content only, and does not depend on the ratio BaO/P_2O_5, but the system containing Na_2O has always lower resistivity than those without Na_2O for the same V_2O_5 content. On the other hand, KITAIGRODSKII and KAYZECHENCO[4] report that, for a constant content of V_2O_5, the resistivity can be reduced by substituting P_2O_5 for PbO, BaO, WO_3 or MoO_3.

As can be seen from the above-mentioned reports, little has been known as to the effect of chemical constituents on the conductivity of electron-conducting glasses such as high-vanadium glass, although their role in the ion-conducting glasses is fairly well understood, at least qualitatively.[5]

The purpose of the present paper is to discuss the relations among the factors determining the conductivity of such glasses, which have been found out during the course of the experiment.

2. EXPERIMENTAL METHODS

V_2O_5, H_3PO_4, $BaCO_3$ and carbonates of other metals of reagent grade have been used for the material. Spectral analysis has shown that the impurities in V_2O_5 are: Fe_2O_3 0·007 per cent and TiO_2 trace. The materials are mixed in a predetermined ratio with some water, dried and, after the reaction with phosphoric acid has sufficiently proceeded, heated gradually in an alumina crucible up to a fixed temperature, kept 30 min there, and are molded into a rod for the measurement of the electrical properties.*

Electrical conductivity was measured by evaporating gold electrodes on both ends of the sample, and applying an electric field of about 1 V/cm. The effect of the resistance between the material and the electrodes was shown to be negligible under these conditions. All the resistivity measurements were performed at 40°C.

3. RESULTS

(1) *The existence of lower-valence vanadium*

In all the samples investigated, the existence of vanadium ion of valence lower than 5 has been detected, the amount being as large as 40 per cent or more of the total vanadium content. When the contents of the glass, namely, H_3PO_4, V_2O_5 and $BaCO_3$, are mixed up and then dissolved in dilute H_2SO_4, the solution does not react at all with $KMnO_4$ solution. However, when the mixture is melted into glass and then dissolved in H_2SO_4, the solution reacts strongly, as a reducing agent, with $KMnO_4$. This reduction must be attributed to lower-valence vanadium ions, since no other reducing ions are likely to exist in the solution. As this effect is only found after the mixture is melted into glassy state, these lower-valence vanadium ions must have been produced while the materials have been melted together into glass.

Whether this lower-valence vanadium is V^{3+} or V^{4+} cannot be determined from this result alone, but the fact that this glass is easily soluble in acid suggests that it is V^{4+}, as it is known that V_2O_4 is easily soluble in acids while V_2O_3 is not. The experiments with samples melted in atmospheres of H_2 or SO_2 mentioned below also confirm this conclusion.

The ratio V^{4+}/V_{total} has been measured for each sample, using the value of V^{4+} calculated from the volume of the $KMnO_4$ solution consumed and the amount of the total vanadium obtained by precipitation with oxine. It has been observed that this ratio decreases if P_2O_5 is substituted for other oxides, keeping the vanadium constant, and increases as the melting temperature is increased.

(2) *Relation between* V^{4+} *content and the resistivity*

The resistivity of the system $60\,V_2O_5 \cdot x\,BaO \cdot (40-x)P_2O_5$† is plotted in Fig. 1, with the melting temperature as a parameter. It is seen that the resistivity varies in a rather complicated way according to the ratio of the contents BaO/P_2O_5 and to the melting temperature. If, however, the resistivity is plotted against V^{4+}/V_{total}, as in Fig. 2, it can be seen that the resistivity is a function of

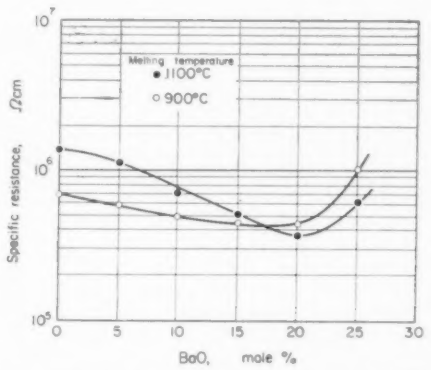


FIG. 1. Change in resistivity of glass containing 60 mole per cent of V_2O_5 as the ratio BaO/P_2O_5 and melting temperature are varied.

* As it has been shown that the resistivity of these glasses does not depend on the thermal history in the transformation range, no heat treatment for stabilization has been performed. For example, the density of the glass 60 V_2O_5 : 10 BaO : 30 P_2O_5 varied from 3·0369 to 3·0053 according to its thermal history, while no variation larger than the experimental uncertainties were observed in their resistivity and activation energy.

† The numbers express the ratio of the contents in mole per cent.

only V^{4+}/V_{total}, regardless of the proportion of the other contents and of the melting temperature.

This fact is also seen when the glass is melted in an atmosphere of SO_2. If a glass $60\ V_2O_5 \cdot 20\ BaO \cdot 20\ P_2O_5$ is melted in an atmosphere of SO_2,

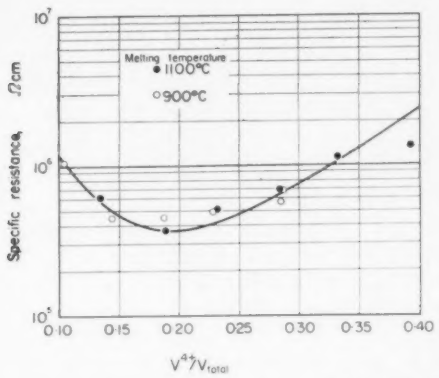


FIG. 2. The same as Fig. 1, plotted against V^{4+}/V_{total}.

its resistivity does not differ appreciably from that of the same glass melted in air; on the other hand, if a glass $60\ V_2O_5 \cdot 5\ BaO \cdot 35\ P_2O_5$ is compared under the same condition, the resistivity of the one melted in an atmosphere of SO_2 increases with the time of melting. This result can be easily understood from Fig. 3, where the resistivity is plotted

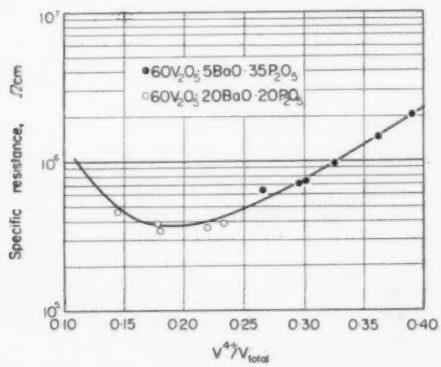


FIG. 3. Relation between resistivity and V^{4+}/V_{total} of two kinds of glass melted in an atmosphere of SO_2.

against V^{4+}/V_{total}. In each case the amount of V^{4+} increases with the time of melting in an atmosphere of SO_2, but in the former case it varies around the value corresponding to the minimum of the resistivity curve, while in the latter case it varies in the region where the curve has a positive slope.

If the glass is melted in an atmosphere of hydrogen, it shows very high resistivity and at the same time it becomes hardly soluble to acids. It is known that V^{3+} is produced when V_2O_5 is heated in hydrogen, while V^{4+} is produced when it is heated in SO_2.[6] Hence it seems quite reasonable that the insolubility of the glass obtained above to acids is due to the production of V^{3+}, and that the lower-valence vanadium appearing in the glass heated in air is V^{4+}.

As the amount of vanadium in the glass is increased, the minimum in the resistivity versus V^{4+}/V_{total} curve corresponding to Figs. 2 or 3 becomes less marked, that is the resistivity becomes less sensible to the ratio BaO/P_2O_5 and to the melting temperature. This would be one of the reasons why the resistivity has been considered to depend on the amount of V only in the previous literatures.

If we plot figures corresponding to Fig. 2 for various vanadium concentration in the system V_2O_5–P_2O_5–BaO and denote with $(V^{4+})_m$ the value of V^{4+} for the minimum of resistivity, and with $(V^{5+})_m$ the value of V^{5+} for the same sample (both in weight per cent), there is a simple empirical relation:

$$\frac{56-(V^{5+})_m}{(V^{4+})_m} \simeq 3{\cdot}0$$

where 56 per cent corresponds to V_2/V_2O_5. The explanation for this relation is not made at present.

(3) *The effect of other ions*

Similar experiments have been performed with BaO substituted for oxides of various kinds of metals of Ia, Ib, IIa or IIb, as well as PbO or Tl_2O, and the resistivity has in each case been measured as a function of V^{4+}/V_{total}. In each case, the resistivity becomes minimum for the value of V^{4+}/V_{total} between 0·1 and 0·2, although the form of the curve and the minimum resistivity, ρ_m, are characteristic of the added ion.

In Fig. 4, the minimum resistivity, ρ_m, is plotted against the "cationic field strength,"[7] Z/a^2, of the added element, where Z is the charge on the cation,

and a is the mean distance between oxygen and each cation in Å.

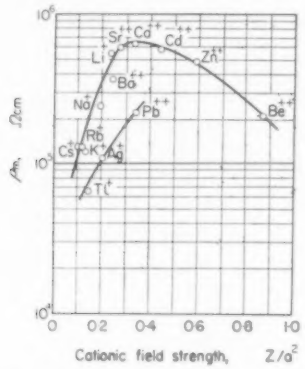


FIG. 4. Relation between minimum resistivity, ρ_m, of the glass $60\ V_2O_5 \cdot x\text{RO} \cdot (40-x)P_2O_5$ and the cationic field strength of R.

The minimum in the resistivity versus V^{4+}/V_{total} curve is not found if transition elements are used for substituting BaO in the glass under question. In these cases, the resistivity decreases or increases monotonically with the quantity of the added element.

4. DISCUSSION

It may be pointed out that there is a strong similarity between the dependence of the resistivity of the glass on V^{4+}/V_{total} mentioned above and the dependence of the resistivity of magnetite on its composition. In the latter case, there is a minimum of resistivity for $Fe^{2+} : Fe^{3+} :: 1 : 2$, and the resistivity increases if the ratio deviates from this value in either direction, although the type of conduction is always n-type. In the latter case, the electrical conduction is explained by the exchange of electrons between neighbouring Fe^{2+} and Fe^{3+} occupying the B-site in the spinel structure.

For the stoichiometric ratio of $Fe^{2+} : Fe^{3+} :: 1 : 2$, the number of neighbouring pairs is maximum and hence resistivity minimum.

From the similarity of the dependence of resistivity on the ratio of the lower-valence ions to the total in both cases, we may conclude that the electrical conduction in our glass is due to the exchange of electrons between lower- and higher-valence vanadium ions through oxygen, since it seems quite reasonable that a mechanism like valence exchange does not necessitate the material to be crystalline.[8] In such a case, the minimum of resistivity may appear when the number of pairs of energetically equivalent V^{4+} and V^{5+} becomes maximum.

Photoconductivity has not been observed in these glasses. Also, the Hall coefficient is very small. Moreover, thermoelectric-power measurement shows that the glass is always n-type regardless of whether the value of V^{4+}/V^{5+} is larger or smaller than the optimum value. These facts would strongly support the assumption about the conduction mechanism in the glass under consideration.

The result shown in Fig. 4 can also be explained with this assumption. As we assume that the electrons are exchanged through oxygen between vanadium ions, the orbit of the valence electrons in oxygen may overlap with the neighbouring vanadium ions, and hence the probability of exchange may depend on the polarization of the oxygen ion, which in turn may depend on the polarizing power of other ions in its neighbourhood. So the resistivity will be smaller if the polarizing power of the neighbouring ion is smaller.

As the cationic field strength in Fig. 4 can be considered as a measure of the polarizing power of each ion in the glass against the oxygen ion, it can easily be understood that its smaller value gives smaller ρ_m value. The fact that the results obtained with ions of b-group lie on a different curve from those obtained with a-group ions, would be due to the difference in the relation between cationic field strength and polarizing power for the two types of ions.

For cationic field strengths of more than 0·35, the increase in this quantity results in the decrease in ρ_m. The explanation is as follows: in vanadium phosphate glass, vanadium and phosphorus act as network formers; other cations exist in this network as network modifiers. Among these, the ions with small cationic field strength can come close to the oxygen ion forming the network V^{4+}–O–V^{5+}, while the binding between oxygen and the ions with stronger cationic field is more covalent,[9] and hence these ions would act rather as network formers than come close to the oxygen ions already forming networks.[10] Thus the electrical conduction of the glass would be less affected by the addition of these ions.

5. CONCLUSIONS

The electrical conduction of semiconducting vanadium glass is affected by the following three factors.

(1) *Vanadium concentration.* As has already been known, other elements, at least other than transition elements, act only as diluting agents, and have nothing to do with the electrical conduction. This is the same as in the case when Fe_3O_4 is diluted with $MgCr_2O_4$, which has the same spinel structure as Fe_3O_4.

(2) *Ratio of* V^{4+} *to the total vanadium.* For the same concentration of vanadium as a total, the resistivity of the glass is minimum for a certain critical value of this ratio, and the resistivity increases as the ratio is varied in either direction. This ratio increases with increasing melting temperature and is larger for larger acidity of the diluent.

(3) *The kind of basic cation.* Among monovalent ions, those with smaller cationic field strength are more effective in lowering the resistivity than those with higher cationic field strength. For larger cationic field strength, namely, larger than 0·35, the effect increases with increasing value of this quantity.

The results can be explained by assuming that the electrical conduction is due to the valence exchange between V^{4+} and V^{5+} through O^{2-}. By choosing suitable values for these factors, it is made possible to obtain semiconducting glass with the room-temperature resistivity of the order of 10^3 Ω-cm.

Acknowledgements—The author wishes to express his sincere thanks to Professor T. MORIYA and Professor S. MAKISHIMA for their interest and suggestions concerning this investigation. Thanks are also due to Dr. G. M. HATOYAMA, Head of Physics Division, Electrotechnical Laboratory, for the encouragements given in the publication of the paper.

REFERENCES

1. E. P. DENTON, H. RAWSON and J. E. STANWORTH, *Nature, Lond.* **173,** 1030 (1954).
2. A. E. DALE and J. E. STANWORTH, *Research, Lond.* **7,** 38 (1954).
3. P. L. BAYNTON, H. RAWSON and J. E. STANWORTH, *Trans. Electrochem. Soc.* **104,** 237 (1957).
4. I. I. KITAIGRODSKII and V. G. KAYZECHENCO, *Glass and Ceramics, Moscow* **15,** (6), 8 (1958); *Ceramic Abstr.* **42** (2), 41 (1959).
5. G. W. MOREY, *Properties of Glass* p. 465. Reinhold, New York (1954).
6. S. TERANISHI, *J. Chem. Phys.* **27,** 1217 (1957).
7. A. DIETZEL, *Glastech. Ber.* **22,** 411 (1948).
8. S. MAKISHIMA. Private communication.
9. H. SALMANG, *Glastech. Ber.* **32,** 137 (1959).
10. A. SMEKAL, *Glastech. Ber.* **22,** 278 (1949).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 164–171. Printed in Great Britain

CALCULATION OF CONCENTRATION PROFILES AND SURFACE CONCENTRATION FROM SHEET-CONDUCTANCE MEASUREMENTS OF DIFFUSED LAYERS*

M. F. LAMORTE†

Westinghouse Electric Corporation, Semiconductor Department, Youngwood, Pennsylvania

(*Received* 18 *August* 1959; *in revised form* 23 *February* 1960)

Abstract—A method for obtaining the sheet-conductance profile of a diffused layer by ohmic contacts to the layer is suggested. Equations are presented from which the concentration profile may be obtained from sheet-conductance measurements by either the four-probe or ohmic-contact method. For three impurity-concentration distributions, equations are presented for surface concentration which necessitate the measurements of sheet conductance, junction penetration and the background conductivity. If the sheet conductance is much greater than the product of the background conductivity and the junction penetration, only the sheet conductance need be measured in order to determine the surface concentration.

Résumé—On suggère une méthode pour l'obtention du profil de la conductivité de couches superficielles d'une couche diffusée par des contacts Ohmiques sur la couche. On donne les équations à partir desquelles on peut obtenir les profils de concentration à l'aide de mesures de la conductivité de couches superficielles soit par la méthode de quatre points, soit par la méthode de contact Ohmique. Pour trois distributions de concentration des impuretés on établit des équations pour la concentration en surface qui nécessite la mesure de la conductivité de couches superficielles de la pénétration de la jonction et de la conductivité en volume. Si la conductivité de couches superficielles est beaucoup plus grande que le produit de la conductivité en volume et la pénétration de jonction, on a seulement besoin de la mesurer de la conductivité de couches superficielles pour déterminer la concentration en surface.

Zusammenfassung—Eine Methode zur Ermittlung der Leitfähigkeit einer diffundierten Schicht mit Hilfe Ohm'scher Kontakte auf der Schicht wird vorgeschlagen. Es werden Gleichungen angegeben, durch die das Konzentrationsprofil aus Messungen der Schichtleitfähigkeit errechnet werden kann und zwar entweder durch die Vierspitzenmethode oder durch Ohm'sche Kontakte. Für drei Verteilungen der Störstellenkonzentration werden Gleichungen für die Oberflächenkonzentration gegeben, die die Messung der Schichleitfähigkeit, Lage des p–n-Überganges sowie Leitfähigkeit des Grundmaterials erfordern. Wenn die Leitfähigkeit der Schicht viel grösser ist als das Produkt aus Leitfähigkeit des Grundmaterials und der Tiefe des p–n-Überganges, braucht nur die Schichtleitfähigkeit gemessen zu werden, um die Oberflächenkonzentration zu bestimmen.

INTRODUCTION

WHILE studying the concentration profiles in semiconductor materials which resulted from solid-state diffusion, we realized that the sheet-conductance method employing the four-probe conductivity apparatus was not sufficiently accurate.[1] Usually, even with great diligence, the measurement is not more accurate than 10 per cent. In the first part of this paper we describe a method and procedure which is superior. The method enables one to obtain concentration profiles and surface-concentration values with greater ease, accuracy and precision. In the second part we present equations in closed form to determine surface concentration. The agreement found among the experimental determinations from concentration-profile extrapolation, the values taken from the curves of BACKENSTOSS,[2] and the results of the equations are quite good for this kind of experiment.

* This work was partly supported by the U.S. Air Force, Dayton, Ohio.

† M. F. LAMORTE is presently employed at Radio Corporation of America, Somerville, New Jersey.

NOTATION

ρ	Resistivity in Ω-cm.
x_j	p–n Junction penetration in cm measured from the surface.
$V(x)$, $I(x)$	Voltage and current, respectively, of either the ohmic contacts or four-probe apparatus in m.k.s. units at the plane denoted by x.
$G_s(x)$	The sheet conductance in $(\Omega/\text{square})^{-1}$ of the layer whose thickness is x_j-x.
$\bar{\sigma}(x)$	The average conductivity of a layer of thickness Δx about x in units of $(\Omega\text{-cm})^{-1}$.
q	$1{\cdot}6\times10^{-19}$ C, the electronic charge in m.k.s. units.
$\overline{N_T}(x)$	The average net impurity concentration in a layer of thickness Δx about x in units cm^{-3}.
$\mu(N)$	The carrier mobility at the concentration N in units cm^2 (V-sec)$^{-1}$.
$G_T(x, x')$	The sheet conductance in $(\Omega/\text{square})^{-1}$ of two diffused skins connected in parallel of thickness x_j-x and $x_j'-x'$.
$\sigma(x)$	The conductivity in units $(\Omega\text{-cm})^{-1}$ at the plane x parallel to the p–n junction plane.
$N_T(x)$	The net impurity concentration in units cm^{-3}.
C_0	The surface concentration in units cm^{-3}.
N_0	The net impurity concentration of the material before diffusion in units cm^{-3}.
$Q_T(x)$	The net impurity surface density between the plane x parallel to the p–n junction and x_j in units cm^{-2}.
μ_0	A carrier mobility which is independent of the concentration in units cm^2 (V-sec)$^{-1}$.
$\bar{\mu}$	An average carrier mobility in units cm^2 (volt-sec)$^{-1}$.
D	The diffusion coefficient in units $\text{cm}^2\,\text{sec}^{-1}$.
t	The diffusion time in units of sec.
u	$x/(2\sqrt{Dt})$.
σ_0	The conductivity of the material before diffusion in units $(\Omega\text{-cm})^{-1}$.
d_{12}	The diffusion coefficient in units of 10^{-12} $\text{cm}^2\,\text{sec}^{-1}$.
h	The diffusion time in units of hr.
$\tilde{\sigma}_{0s}$	$C_0\bar{\mu}q$.
v	The velocity of a receding surface which is evaporating in units cm sec^{-1}.
X	$x-vt$.
σ_{eff}	$G_s(0)/x_j$, the effective conductivity of a diffused skin.
S	The distance in cm between the point probes of the four-probe apparatus.

SHEET-CONDUCTANCE PROFILES

We shall confine our attention to the sheet-conductance method to determine profiles. Prior to diffusing, samples from which profiles are desired are polished with Linde "B" powder on a slow-speed polishing wheel to a very high gloss. Polishing with Linde "B" powder (particle size 0·1–0·3 μ) insures that the disturbed layer on a silicon surface is less than 0·3 μ. In the diffusion furnace the sample is supported on a flat quartz platform in order to prevent any warping of the polished surfaces. After diffusion, the specimen is mounted on a gauge block using a hard cement and making certain that a thin layer is present between the sample and the block. The sides of the gauge block are placed into position against three drill rods mounted vertically in the base of a Zeiss optimeter to obtain the penetration distance. The drill rods insure that the optimeter measurement is always taken in the same region. In the region in which the optimeter measurement is recorded, the four-probe measurement is also taken. Once the experiment is started, the sample is not removed or its position changed. After first polishing with Linde "B" powder, hydrofluoric acid and deionized water rinses are applied, and the electrical measurements, followed by the optimeter measurements, are then taken. This procedure is repeated until the diffused skin has been removed. Sheet conductance vs. penetration distance is then plotted on semi-logarithmic graph paper. A smooth curve is drawn through the data, and it is from this curve that the profile is calculated.[1]

The polishing is accomplished by hand or by using a precision polisher.[3] Between measurements the sample is polished for approximately 1–2 min removing between 0·01 and 0·02 mil. Care must be excerised while polishing to avoid producing a disturbed layer or scratches, which result in a highly conductive layer. To determine whether equal amounts are removed from different regions of the surface, several optimeter measurements are taken.

The equation for the four-probe resistivity with equally spaced collinear array for a diffused skin is[4]

$$\rho = 4{\cdot}53x_jV/I \qquad (1)$$

under the condition that $x_j \ll \frac{1}{2}S$. The factor 4·53 represents the numeric $\pi/\ln 2$. The sheet conductance G_s may be expressed in units of $(\Omega/\text{square})^{-1}$. At the original surface, $G_s(0) = x_j/\bar{\rho}$. The change in G_s resulting from removing a layer Δx from the surface at x is given by

$$\Delta G_s(x) = (1/4{\cdot}5)\{I(x)/V(x) - I(x+\Delta x)/V(x+\Delta x)\}$$

In terms of semiconductor parameters of the diffused skin this change is given by $\bar{\sigma}(x)\Delta x$, where

$\bar{\sigma}(x)$ is the average value of the conductivity in the layer Δx. Therefore, the change in sheet conductance may be written as $\bar{N}_T(x)q\mu(\bar{N}_T)\Delta x$ where $\bar{N}_T(x)$ is the average net impurity concentration in the layer Δx. We then solve for this concentration, resulting in

$$\bar{N}_T(x) = \frac{1}{4{\cdot}5q\mu(\bar{N}_T)\Delta x}\left[\frac{I(x)}{V(x)} - \frac{I(x+\Delta x)}{V(x+\Delta x)}\right] \quad (2)$$

from which the net impurity-concentration profile may be obtained.

In addition to the inherent 10 per cent error, the four-probe apparatus has several disadvantages. Frequent pointing or replacement of the probes is necessary to make good electrical contact. For reliable measurements it is sometimes necessary to sandblast the surface prior to placing the probes on the surface.

Because the reliability of the apparatus is cast into doubt in the high-resistivity ranges for both Ge and Si, other methods have been explored. A method suggested here is to obtain the sheet-conductance profile by means of ohmic contacts to the diffused skin placed at the ends of a thin filament which has been diffused. The filament is prepared identically to that used for the four-probe method, except that in this case ohmic contacts are fused onto the diffused skin. Sheet conductance is then given as the product of the ratio of the current through the ohmic contacts to the voltage between the ohmic contacts and the ratio of the length to the width of the filament. Typical filament dimensions are $10\times 20\times 1$ mm. The relation for the average net impurity concentration in the layer Δx for this case is

$$\bar{N}_T(x) = \frac{1}{q\mu(\bar{N}_T)\Delta x}\left[\frac{I(x)}{V(x)} - \frac{I(x+\Delta x)}{V(x+\Delta x)}\right] \quad (3)$$

It is to be noted that equation (3), for the ohmic-contact method, differs by the factor 4·5 from equation (2), for the four-probe method. It is felt that in the ohmic-contact method the electrical measurements are more precise provided that the contact potential drop across the ohmic contact is negligible compared to the measured voltage. Clearly, the layer removed from the surface during polishing must be of uniform thickness.

If both surfaces of a filament have been polished and diffused and if the diffused skin along the sides has then been removed by polishing or etching, the total sheet conductance measured is

$$G_T(x, x') = G(x)+G(x') \quad (4)$$

where $G(x)$ and $G(x')$ are the contributions to the measured sheet conductance of the top and bottom sides of the filament, respectively, as it is mounted on the gauge block. By symmetry, $G(x = 0) = G(x' = 0)$, whence

$$G_T(0, 0) = 2G(0)$$

In terms of measured quantities the sheet conductance of the top surface in the presence of a diffused skin on the bottom surface is

$$G(x) = G_T(x, 0)-G_T(0, 0)/2 \quad (5)$$

The quantity $G(x)$ is plotted to obtain the sheet-conductance profile rather than the measured value $G_T(x, 0)$. It is seen that the maximum change that $G_T(x, 0)$ can undergo by polishing the top surface is $G_T(0, 0)/2$.

Fig. 1 exhibits the sheet conductance measured

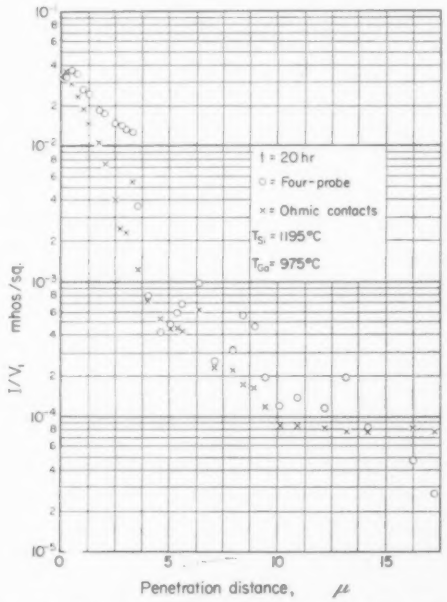


FIG. 1. Sealed-tube vacuum gallium diffusion. Conductance vs. penetration distance. The data as recorded for the four-probe and ohmic-contact methods are shown for the same gallium-diffused skin.

on the same gallium diffused sample employing the four-probe apparatus and by means of ohmic contacts attached to the skin. It is evident that the four-probe data exhibit a large spread while those for ohmic contacts, in comparison, show much less. In the range of conductance values less than 10^{-2} mho the spread in the four-probe data is particularly bad; however, this is not surprising for it is well known that the four-probe apparatus is inaccurate for measuring resistivities above 200 Ω-cm. This is typical of data obtained in this range of conductance.

When employing the ohmic-contact method the junction leakage is determined after the diffused skin has been removed. If the periphery around the p–n junction has been etched and masked when mounted on the gauge block, the leakage is usually constant. This leakage conductance is subtracted from the data. The resultant curve is the one from which the profile is calculated. In Fig. 1 we see that this conductance is approximately 8×10^{-5} mho.

Consider a square portion of a diffused skin shown in Fig. 2. If the impurity concentration is a

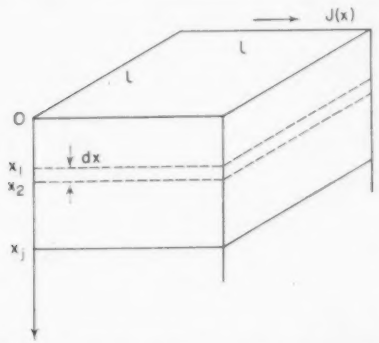


FIG. 2. A square portion of a diffused layer exhibiting the p–n junction at X_j.

function of x, so is the conductivity, though it does not necessarily have the same functional dependence. The current density parallel to the surface produced by an electrical-field intensity E is also a function of x. The sheet conductance in differential form is obtained:

$$dG_s(x) = dI(x)/V = N_T(x)q\mu(N_T)\,dx \qquad (6)$$

To determine the current parallel to the surface within the diffused skin, one integrates equation (6) from $x = 0$ to $x = x_j$. The sheet conductance of a diffused skin becomes[2]

$$G_s(0) = q\int_0^{x_j} N_T(x)\mu(N_T)\,dx \qquad (7)$$

In that which follows, equations are presented for the ohmic-contact method along with the four-probe, and all equations are on a per-unit-area basis. The differential sheet conductance may then be written as

$$dG_s(x) = -d\{I(x)/V(x)\}, \text{ ohmic contacts} \qquad (8a)$$

$$dG_s(x) = -(1/4{\cdot}5)\,d\{I(x)/V(x)\}, \text{ four-probe} \qquad (8b)$$

Because $dG_s(x)$ is a positive quantity, and since the slope of the curve $I(x)/V(x)$ vs. x is negative, a minus sign is present on the right-hand side of equations (8).

It is interesting to note that sheet conductance is independent of the impurity-concentration distribution and junction penetration only when the following conditions apply: (1) that the total impurity atoms per unit surface area remains constant; (2) that the mobility is constant and independent of the impurity concentration. We define

$$Q_T(x) = \int_x^{x_j} N_T(x)\,dx \qquad (9)$$

as the total impurity atoms per cm^2 of surface area between two parallel planes whose co-ordinates are x and x_j. In the above, we take $Q_T(0) = $ a constant. Under the above conditions $G_s(0) = q\mu_0 Q_T(0)$, where μ_0 is the constant mobility taken independent of concentration, and $G_s(0)$ is constant as we wanted to show.

Generally, however, when the mobility is a function of the impurity concentration, the sheet conductance is then a function of the concentration distribution and the functional dependence of mobility on concentration. Integrating $G_s(x)$ by parts results in

$$G_s(x) = q\mu(N_T)Q_T(x) + q\int_x^{x_j} Q_T(x)\,d\mu(N_T) \qquad (10)$$

Where mobility is independent of concentration

the integral vanishes and the previous result that $G_s(0)$ is independent of the distribution is recovered. Moreover, we find that $G_s(x_j)$ vanishes as expected. For those distributions where the net total atoms per cm^2 of surface area is large, the function is dominated by $Q_T(x)$. In those distributions where $C_0 > N_0$ and $N_T(x)$ decreases rapidly, the region nearest to the surface contributes most to the sheet conductance. This is true because the contribution to $Q_T(x)$ is greatest in the surface region. Moreover, for $N_0 < 10^{16}$ the integral may be cut off for some $x < x_j$ since in this region $d\mu(N_T)$ vanishes. In short, such cases result in the integral being less sensitive to the value of x_j.

SPECIFIC CONCENTRATION DISTRIBUTIONS

The general equations up to this point do not involve specific distributions. In order to investigate specific distributions one of two approximations is employed. The first involves the assumption of an average mobility $\bar{\mu}$ throughout the diffused skin in those integrals for which the integrand is $N_T(x)\mu(N_T)$. The average mobility is determined from the relation

$$\bar{\mu} = \int_0^{x_j} N_T(x)\mu(N_T)\,dx \left[\int_0^{x_j} N_T(x)\,dx\right]^{-1} \qquad (11)$$

It is a poor approximation in some instances while good in others. It becomes better as the quantity $C_0 - N_0$ decreases. When there is no variation in concentration (i.e. $C_0 - N_0$ = a constant or vanishes) the approximation becomes exact.

A better approximation is suggested. The exact expression for $\sigma(x)/q$ is

$$N_T(x)\mu(N_T) = N(x)\mu(N_T) - N_0\mu(N_T)$$

The approximation used here is to replace $\mu(N_T)$ by $\mu(N)$ in the first term and $\mu(N_T)$ by $\mu(N_0)$ in the second, resulting in

$$N_T(x)\mu(N_T) = N(x)\mu(N) - N_0\mu(N_T) \qquad (12)$$

In regions where $N(x) \gg N_0$ the term $N_0\mu(N_0)$ is negligible and the approximation is good. As $N(x)$ approaches N_0 the three mobilities $\mu(N_T)$, $\mu(N)$ and $\mu(N_0)$ are very nearly equal. In Fig. 3 the exact equation and the approximation are plotted for several cases for n- and p-type layers. It is seen that the approximation is a very good one.

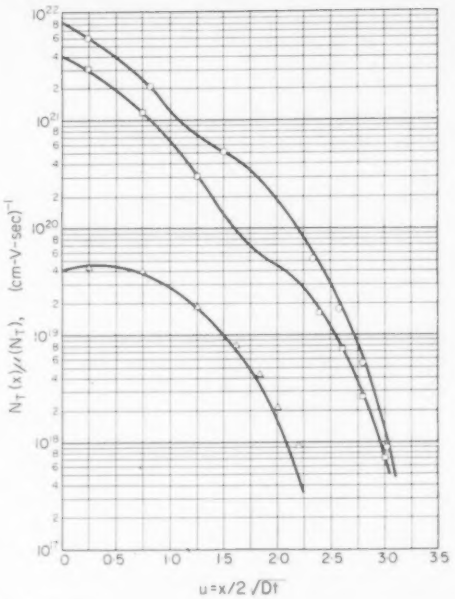


FIG. 3. Comparison of the exact concentration–mobility product with the approximate value defined in equation (12), vs. u. Using the mobility data of BACKENSTOSS[5] and the concentration distribution C_0 erfc$u$, the upper curve is calculated for n-type, $C_0 = 10^{20}$ and $N_0 = 10^{15}$; the middle curve for p-type, $C_0 = 10^{20}$, $N_0 = 10^{15}$; the lower for p-type, $C_0 = 10^{18}$ and $N_0 = 10^{15}$. The circles, squares and triangles represent the approximations.

In conjunction with the first term on the right-hand side of equation (12), it is shown in Fig. 4 that over a considerable range $N(x)\mu(N)$ exhibits a simple functional dependence on $N(x)$. The data of Fig. 4 were obtained from those for conductivity mobilities for both n- and p-type silicon presented by BACKENSTOSS.[5] The equations are of the form

$$N(x)\mu(N) = 10^B N(x)^A \qquad (13)$$

For n-type silicon, $B = 8{\cdot}375$, $A = 0{\cdot}675$; while $B = 7{\cdot}169$, $A = 0{\cdot}704$ for p-type. The equations in this concentration range are:

$$N(x)\mu(N) = 2{\cdot}37 \times 10^8 \times N(x)^{0{\cdot}675},\ n\text{-type Si} \qquad (14a)$$

$$N(x)\mu(N) = 1{\cdot}47 \times 10^7 \times N(x)^{0{\cdot}704},\ p\text{-type Si} \qquad (14b)$$

Wherever possible these relations are employed in the integrands of our equations. In those regions where the mobility is relatively constant and the functional dependence of equation (13) is not

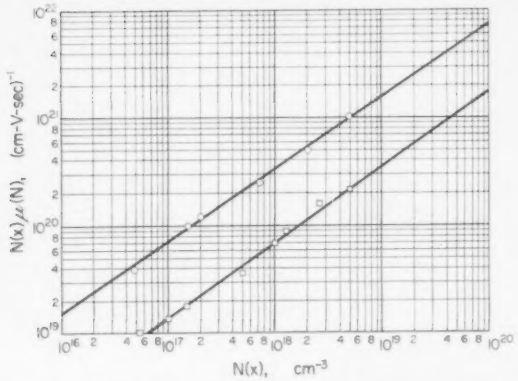


FIG. 4. $N(x)\mu(N)$ vs. $N(x)$, from the data of BACKENSTOSS[5].

obeyed, the integrals may be broken off to insert this condition.

We will investigate several of the more common distributions used at the present time. We will obtain analytic solutions for surface concentration and sheet conductance in closed form for the distributions. Recently a set of curves was published relating surface concentration to the product of the measured sheet conductance, $G_s(0)$, and junction penetration.[2] Prior to that C_0 was determined by chemical analysis, extrapolating the concentration profile to the surface,[1] or by the junction-penetration method.[6] In addition to the junction penetration, the equations derived here for C_0 require that the diffusion coefficient, the thermally stabilized background resistivity and $G_s(0)$ be known, and a specific distribution assumed. The solutions in closed form are expressed in terms of these parameters.

I. *Complementary error function*: $N_T(x) = C_0 \operatorname{erfc} u - N_0$

Substituting the concentration distribution and the approximation given by equation (12) in that for $G_s(x)$, we obtain

$$G_s(x) = 2\sqrt{(Dt)}q \int_u^{u_j} \mu(N) C_0 \operatorname{erfc} u'\, du' - \\ - qN_0\mu(N_0)(x_j - x) \qquad (15)$$

The integration of $\mu(N)C_0 \operatorname{erfc} u'\, du'$ is difficult to perform unless it is expanded into a converging series. To obtain a convenient expression we also make use of the approximation presented in equation (11), resulting in

$$G_s(x) = 2\sqrt{(Dt)}\bar{\sigma}_{0s} \int_u^{u_j} \operatorname{erfc} u'\, du' - (x_j - x)\sigma_0 \qquad (16)$$

where σ_0 is the thermally stabilized background conductivity. When x is permitted to go to zero we obtain C_0 in terms of quantities which may be measured.

$$C_0 = \tfrac{1}{2}[G_s(0) + x_j\sigma_0]\left[\sqrt{(Dt)}\bar{\mu}q \int_u^{u_j} \operatorname{erfc} u'\, du'\right]^{-1} \qquad (17)$$

The two equations for $I(x)/V(x)$ follow from equation (16). However, when $u_j > 1{\cdot}6$ a simplification results in equation (17) giving

$$C_0 = \{G_s(0) + x_j\sigma_0\}/\{1{\cdot}13\sqrt{(Dt)}\bar{\mu}q\} \qquad (18)$$

since in this range

$$\int_0^{u_j} \operatorname{erfc} u'\, du' \cong \int_0^{\infty} \operatorname{erfc} u'\, du' = 1{\cdot}13/2$$

to within 1 per cent. Also, for $u_j > 1{\cdot}6$ we have that $x_j > 1{\cdot}92\sqrt{(d_{12}h)}$ in units of microns.

II. *Steady-state distribution*

$N(x) = C_0 \exp(-vX'/D),\ X' = x - vt$

The analytic expression for $N(x)\mu(N)$ may be employed for this distribution in the appropriate concentration range in conjunction with the approximation given in (12). Accordingly, substituting into $G_s(X)$ we obtain

$$G_s(X') = [qD \times 10^B/Av]C_0A \times \\ \times [\exp(-AvX'/D) - \exp(-AvX_j'/D)] - \\ -(X_j' - X')\sigma_0 \qquad (19)$$

Permitting X' to vanish we then solve for C_0:

$$C_0 = [\{G_s(0) + X_j'\sigma_0\}\{qD \times 10^B/Av\}^{-1} \times \\ \times \{1 - \exp(-AvX_j'/D)\}^{-1}]^{1/A} \qquad (20)$$

III. *Gaussian curve distribution*

$N(x) = Q/\sqrt{(\pi Dt)} \exp(-x^2/4Dt)$

Again we may make use of the analytic functions

for $N(x)\mu(N)$ in conjunction with the approximation given in equation (13). For convenience, we define $y = x\sqrt{(A/4Dt)}$. The result of substituting this value into $G_s(x)$ is

$$G_s(x) = q\times 10^B C_0{}^A\sqrt{(\pi Dt/A)}\times \times[\operatorname{erfc} y_j - \operatorname{erf} y] - (x_j - x)\sigma_0 \quad (21)$$

making use of

$$\int_y^{y_j} e^{-\beta^2}\,d\beta = \sqrt{(\pi)}/2[\operatorname{erf} y_j - \operatorname{erf} y]$$

where

$$\operatorname{erf} y = 2/\sqrt{(\pi)}\int_0^y e^{-\beta^2}\,d\beta$$

When $x = 0$ we then solve for C_0:

$$C_0 = [\{G_s(0)+x_j\sigma_0\}/\{q10^B\sqrt{(\pi Dt/A)}\operatorname{erf} y_j\}]^{1/A} \quad (22)$$

The rate of change of erfc u with respect to x decreases monotonically as x increases. However, the linear term in $G_s(x)$ exhibits a constant rate of change which is positive and equal to σ_0. But $d\{I(x)/V(x)\}/dx$ is always negative, since the sheet conductance is a monotonically decreasing function. We infer from this that the rate of change of the first term with respect to x is negative, but greater in magnitude than σ_0 for all $x < x_j$. This interpretation becomes more lucid when N_0 is substituted for $N_T(x)$ in $G_s(x)$, resulting in $(x_j - x)\sigma_0$. In this case, the change of sheet conductance is attributed only to the fact that the skin thickness decreases: for equal increments of x, a constant surface density of impurity atoms is removed from the layer. When an impurity-concentration distribution is also present, the sheet-conductance rate of change is modified by the change of the rate of removal of impurity atoms for equal increments of x. At the junction, therefore, the derivative of the sheet conductance with respect to x vanishes, resulting in

$$\operatorname{erfc} u_j = \sigma_0/\bar{\sigma}_{0s}$$

But this is the relation resulting from the condition that $N(x_j) = N_0$ which is employed to determine C_0 and D.[6]

It appears from the expressions for C_0 that the equations are not in the form to give results which agree with the published curves.[2] Let us recall that for the complementary-error function and Gaussian-curve distributions x_j is proportional to $\sqrt{(Dt)}$; and X_j is proportional to D/v when the steady-state distribution is present. Defining $\sigma_{\text{eff}} = G_s(0)/x_j$ the numerators of the expressions for C_0 may be placed in the form $\sigma_{\text{eff}}+\sigma_0$, while the denominators are constants depending on x_j. A linear relationship is then exhibited between log C_0 and logs$[\sigma_{\text{eff}}+\sigma_0]$. This relationship is in agreement with the published curves. When $\sigma_{\text{eff}} \gg \sigma_0$ a linear relationship exists between log C_0 and log σ_{eff}. As σ_{eff} approaches the order of σ_0 the curve log C_0 vs. log σ_{eff} exhibits a decreasing slope. Obviously σ_{eff} can never be smaller than σ_0. The smallest value σ_{eff} can assume is σ_0.

In the expressions for C_0, the junction penetration is involved in the limits of an integral and a linear term, rather than in an exponential. As a result C_0 in these expressions is not as sensitive to the junction penetration as the earlier method described.[6] For deep penetrations or high surface concentrations, the expression is even less sensitive to this measurement.

The equation for C_0 for the steady-state distribution was checked to determine its accuracy. Gallium was diffused into an n-type filament in an evacuated sealed quartz tube. Concentration profiles were obtained by means of ohmic contacts from which C_0 is estimated. Assuming that we have the steady-state distribution, the velocity of Si evaporation is obtained from the slope of the straight-line approximation to the concentration profile. Surface concentration is calculated from equation (20). The results of the estimation from the profile, the calculation and that obtained from the curve of BACKENSTOSS[2] are presented in Table 1. From the tabulation it is seen that the values of surface concentration calculated from equation (20) show better agreement with the experimental values than do the values from the curve of Fig. 8 of BACKENSTOSS.

We note from equation (20) that the velocity of evaporation is necessary to calculate C_0. In the table we have obtained v from the slope of the concentration profile. Consequently, we have obtained an additional piece of information from the experimental data. From the distribution function

Table 1

Extrapolated C_0	v (Å/sec)	Calculated C_0 (equation 20)	Curve of BACKENSTOSS
3×10^{17}	1·2	$3{\cdot}5\times10^{17}$	$1{\cdot}5\times10^{17}$
1×10^{17}	1·6	$1{\cdot}2\times10^{17}$	$4{\cdot}0\times10^{16}$
1×10^{19}	3·4	$3{\cdot}4\times10^{18}$	$5{\cdot}0\times10^{17}$
8×10^{19}	3·6	$8{\cdot}8\times10^{19}$	$1{\cdot}5\times10^{19}$
5×10^{16}	2·7	$1{\cdot}4\times10^{17}$	$1{\cdot}5\times10^{17}$

it is seen that the velocity of evaporation is proportional to the slope of the fractional change of the concentration with respect to X. As a result the agreement between the experimentally determined value and the value obtained from equation (20) is better than that obtained from the curve of BACKENSTOSS. In effect, in equation (20) we are substituting a quantity involving the concentration and another quantity involving the first derivative. In studying profiles of a large number of samples it has been found that profiles may vary considerably from theoretical curves. When the profile deviates appreciably from the theoretical curves then the information $G_s(0)$, X_j and σ_0 is not adequate to determine C_0.

The poor agreement between the extrapolated value and that from equation (20) of No. 3 sample is due to the large slope for which the extrapolation was poor.

We have undertaken to show correspondence for one distribution only. The other two distributions should show better agreement between the values of the last two columns since the slope of the profile is not needed in equation (20) for these.

CONCLUSIONS

Whereas obtaining the measurements required to calculate the concentration profile of a diffused skin is time consuming, the profile provides the means by which anomalies may be detected. The exact equations should be employed to calculate the concentration profile from the sheet-conductance profile. A method is suggested for obtaining sheet-conductance profiles of the diffused skin by means of ohmic contacts to the skin, as an alternative to the four-probe method. Sheet conductances of the order of 10^{-2} $(\Omega/\text{square})^{-1}$ require that the ohmic-contact method be used to obtain the profile. The four-probe method cannot be employed in this range because its precision is poor. For greatest precision, the ohmic-contact method is to be employed in all ranges. Equations are derived for three concentration distributions from which the surface concentration may be calculated, and the steady-state distribution is compared with values obtained from profiles and with the curve presented by BACKENSTOSS.

REFERENCES

1. C. S. FULLER and J. A. DITZENBERGER, *J. Appl. Phys.* **27,** 544, (1956).
2. G. BACKENSTOSS, *Bell Syst. Tech. J.* **37,** 699 (1958).
3. W. C. DUNLAP, JR., *Phys. Rev.* **94,** 1531 (1954).
4. A. UHLIR, JR., *Bell Syst. Tech. J.* **34,** 105 (1955).
5. G. BACKENSTOSS, *Phys. Rev.* **108,** 1416 (1957).
6. C. S. FULLER, *Phys. Rev.* **86,** 136 (1952).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 172–175. Printed in Great Britain

DER HALLEFFEKT IM INHOMOGENEN MAGNETFELD

JULIUS BRUNNER

Aus dem Laboratorium der Zentralwerksverwaltung der Siemens-Schuckertwerke AG, Nürnberg, Germany

(*Received* 26 *February* 1960)

Zusammenfassung—Die Hallspannung eines langgestreckten Hallgenerators in einem inhomogenen Magnetfeld, dessen Normalkomponente bezüglich des Hallgenerators in Längsrichtung des Hallgenerators linear verläuft, wird berechnet. Für einen über die Fläche des Hallgenerators weitgehend konstanten Gradienten der Normalkomponente wird daraus der Fehler bei der Feldmessung berechnet und nachgewiesen, dass er in den meisten praktisch vorkommenden Fällen vernachlässigt werden kann.

Abstract—The Hall voltage of a Hall generator with a large ratio of length to width in an inhomogeneous magnetic field is calculated for the case when the normal field component varies linearly along the length of the generator. The error in measurement of fields is shown to be negligible in most cases of practical importance if the gradient of the normal component across the generator is approximately constant.

Résumé—On calcule la tension de Hall d'un générateur de Hall longitudinal dans un champ magnétique dont la composante normale varie linéairement dans la direction de la longueur du générateur de Hall. Dans ce cas on calcule l'erreur sur la mesure du champ pour un gradient constant de la composante normale le long de la surface du générateur de Hall et on montre que cette erreur est négligeable.

1. EINLEITUNG

MAGNETFELDER lassen sich mit Hilfe des Halleffekts nachweisen und messen. Elemente, mit denen man unter der Einwirkung eines Magnetfelds eine durch den Halleffekt bewirkte Hallspannung erzeugen kann, nennt man Hallgeneratoren[1,2].

Wird ein Hallgenerator von einem konstanten Steuerstrom i_s durchflossen und senkrecht von einem homogenen Magnetfeld B durchsetzt (Abb. 1), dann entsteht an den beiden Hallelektroden eine Leerlaufhallspannung

$$u_{h0} = -\frac{R_h}{d} G i_s B \qquad (1)$$

Hierin ist R_h die Hallkonstante des Halbleitermaterials (für Messzwecke vornehmlich InAs bzw. InAsP), d die Dicke der Halbleiterschicht und G eine Grösse, die von der Geometrie des Hallgenerators und gegebenenfalls noch vom Magnetfeld B abhängt[3]. Bei einem langgestreckten Hallgenerator, d.h. einem Seitenverhältnis Länge a/Breite $b \gg 1$, und punktförmigen Hallelektroden ist die Geometriefunktion G gleich 1[3] und somit nach Gleichung (1) die Leerlaufhallspannung u_{h0} der magnetischen Induktion B direkt proportional.

Gleichung (1) gilt nur unter der Voraussetzung, dass die magnetische Induktion B über die Fläche

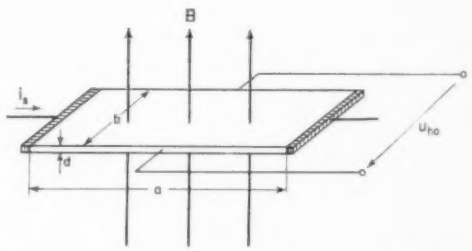


ABB. 1. Stromdurchflossener Hallgenerator im Magnetfeld B.

des Hallgenerators konstant ist. Aus der in einem inhomogenen Feld entstehenden Hallspannung können nur dann Rückschlüsse auf das Magnetfeld gezogen werden, wenn der Zusammenhang zwischen den das Feld bestimmenden Grössen und der Hallspannung bekannt ist. Im folgenden wird

dieser Zusammenhang berechnet für ein Magnetfeld, dessen z-Komponente von x linear abhängt

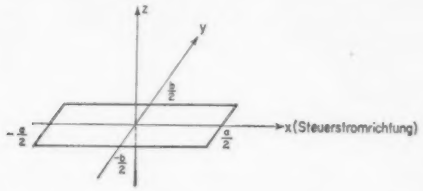


ABB. 2. Zur Wahl des Koordinatensystems.

(vgl. Abb. 2). Die Folgerungen, die sich daraus für die Feldmessung ergeben, werden diskutiert.

2. DIE HALLSPANNUNG EINES HALLGENERATORS IM MAGNETFELD MIT LINEARER x-ABHÄNGIGKEIT DER z-KOMPONENTE

2.1. *Voraussetzungen*

Für die anschliessend durchgeführten Rechnungen werden zwei wesentliche Voraussetzungen gemacht:

(a) Der Hallgenerator soll ein Seitenverhältnis $a/b \gg 1$ haben.

(b) Die Dicke d der Halbleiterschicht soll so klein sein, dass die Abhängigkeit des Magnetfeldes von z (vgl. Abb. 2) im Bereich $-d/2 \leqslant z \leqslant d/2$ vernachlässigt werden kann.

2.2. *Aufstellung der Grundgleichungen*

Wegen der unter 2.1 gemachten Voraussetzung (b) können wir die Grundgleichungen zweidimensional in der x–y-Ebene formulieren. Für die Flächenstromdichte

$$\boldsymbol{J} = d\boldsymbol{j} \tag{2}$$

gilt das um das Lorentz–Glied erweiterte Ohmsche Gesetz

$$J_x = \sigma\, dE_x + \sigma R_h B_z J_y \tag{3a}$$

$$J_y = \sigma\, dE_y - \sigma R_h B_z J_x \tag{3b}$$

σ ist die Leitfähigkeit des Halbleitermaterials. Ferner muss das Stromdichtefeld $\boldsymbol{J}$ quellenfrei und das elektrische Feld $\boldsymbol{E}$ wirbelfrei, d.h.

$$\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} = 0 \tag{4}$$

und

$$\frac{\partial E_x}{\partial} - \frac{\partial F_y}{\partial x} = 0 \text{ sein.} \tag{5}$$

Auf Grund der Voraussetzung (a) in 2.1 haben wir für das Stromdichtefeld $\boldsymbol{J}$ als einzige Randbedingung

$$J_y = 0 \text{ auf den Rändern } y = \pm\frac{b}{2} \tag{6}$$

zu erfüllen.

2.3. *Lösung der Grundgleichungen*

Aus den Gleichungen (3) bis (5) folgt

$$\frac{\partial J_x}{\partial y} - \frac{\partial J_y}{\partial x} = \sigma R_h\left(J_x\frac{\partial B_z}{\partial x} + J_y\frac{\partial B_z}{\partial y}\right) \tag{7}$$

Das Stromdichtefeld $\boldsymbol{J}$ ist durch die Gleichungen (4) und (7) und die Bedingung (6) bestimmt. Die einfachste Lösung, die auch die Randbedingung (6) erfüllt, erhält man mit

$$J_y \equiv 0 \tag{8}$$

Wegen (4) ist dann

$$J_x = J_x(y) \tag{9}$$

also nur von y abhängig. Mit (8) und (9) folgt aus (7):

$$\frac{dJ_x(y)}{dy}\cdot\frac{1}{J_x(y)} = \sigma R_h\frac{\partial B_z}{\partial x} \tag{10}$$

Weil die linke Seite von (10) nur von y abhängt, muss $\partial B_z/\partial x$ von x unabhängig sein, d.h.

$$B_z = f(y)x + g(y) \tag{11}$$

Umgekehrt müssen alle Magnetfelder von der Form (11) zu dem durch (8) und (10) beschriebenen Stromdichtefeld führen, weil nur eine einzige Lösung der Gleichungen (4) und (7) existieren kann, die auch (6) erfüllt.

Mit $\partial B_z/\partial x = f(y)$ ergibt sich als Lösung von (10)

$$J_x = C\exp(\sigma R_h\int_0^y f(y)\,dy) \tag{12}$$

Die Integrationskonstante C ist gemäss

$$\int_{-b/2}^{b/2} J_x(y)\,dy = i_s \tag{13}$$

durch den Steuerstrom i_s bestimmt.

2.4. *Berechnung der Hallspannung*

Die zwischen den Hallelektroden auftretende

Leerlaufhallspannung u_{h0} ist das Integral über die negative y-Komponente der elektrischen Feldstärke entlang der y-Achse von $-b/2$ bis $+b/2$, also

$$u_{h0} = -\int_{-b/2}^{b/2} (E_y)_{x=0}\, dy \tag{14}$$

Mit (3b), (8), (11) und (12) erhält man

$$u_{h0} = -\frac{R_h}{d} C \int_{-b/2}^{b/2} g(y)\, dy \exp(\sigma R_h) \int_0^y f(y)\, dy \tag{15}$$

Für das in (11) enthaltene spezielle Magnetfeld

$$B_z = B_0 + a_1 x + a_2 y \tag{16}$$

(B_0, a_1 und a_2 sind Konstanten.) ergibt sich aus (15) mit (13):

$$u_{h0} = -\frac{R_h}{d} i_s \left\{ B_0 + \frac{a_2 b}{2} \left(\operatorname{ctgh} \frac{\sigma R_h b a_1}{2} - \frac{2}{\sigma R_h b a_1} \right) \right\} \tag{17}$$

3. FOLGERUNGEN FÜR DIE MESSUNG VON MAGNETFELDERN

Gleichung (15) zeigt, dass in einem beliebigen Magnetfeld, dessen z-Komponente nur von y abhängt (d.h. $f(y) \equiv 0$), die Hallspannung dem Mittelwert der z-Komponente über die Breite b des Hallgenerators proportional ist.

Im speziellen Magnetfeld (16), bei dem grad B_z konstant ist, hängt nach (17) die Hallspannung nicht mehr allein von der magnetischen Induktion B_0 im Mittelpunkt $x = y = 0$ des Hallgenerators ab, sondern es tritt zu B_0 ein Korrekturglied, das die partiellen Ableitungen $a_1 = \partial B/\partial x$ und $a_2 = \partial B/\partial y$ enthält. Um den Einfluss des Korrekturgliedes auf die Messung von Magnetfeldern mit

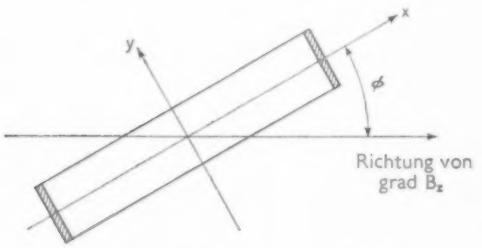


ABB. 3. Zur Definition des Drehwinkels φ.

konstantem Gradienten der z-Komponente bequem diskutieren zu können, führen wir in Gleichung (17)

$$a_1 = |\operatorname{grad} B_z| \cos\varphi, \tag{18}$$

$$a_2 = -|\operatorname{grad} B_z| \sin\varphi \tag{19}$$

ein mit φ als Winkel zwischen der Richtung von grad B_z und der x-Achse des Hallgenerators (vgl. Abb. 3). Man erhält

$$u_{h0} = -\frac{R_h}{d} i_s \left[B_0 - \frac{b}{2} |\operatorname{grad} B_z| \sin\varphi \times \right.$$
$$\times \left\{ \operatorname{ctgh}\left(\frac{\sigma R_h b}{2} |\operatorname{grad} B_z| \cos\varphi \right) - \right.$$
$$\left. \left. - \frac{2}{\sigma R_h b |\operatorname{grad} B_z| \cos\varphi} \right\} \right] \tag{20}$$

Unter der Voraussetzung

$$|\operatorname{grad} B_z| \leqslant \frac{1}{\sigma R_h b} \tag{21}$$

lässt sich Gleichung (20) entwickeln, und es ergibt sich

$$u_{h0} = -\frac{R_h}{d} i_s \left\{ B_0 - \frac{\sigma R_h b^2}{24} |\operatorname{grad} B_z|^2 \sin 2\varphi \right\} \tag{22}$$

Die durch Gleichung (21) gegebene Einschränkung des Gültigkeitsbereiches von Gleichung (22) ist in den meisten praktischen Fällen ohne Bedeutung. Für Hallgeneratoren aus InAs mit $\sigma R_h \sim 20\,000$ $\mathrm{cm^2 V^{-1} sec^{-1}}$ und $b \sim 0{,}5$ cm gilt Gleichung (21) und damit auch Gleichung (22) bis zu Gradienten von 10 000 Gauss $\mathrm{cm^{-1}}$. Durch die Verwendung von Hallgeneratoren mit kleinerer Breite b und Halbleitermaterialien geringerer Beweglichkeit (z.B. InAsP) kann der Gültigkeitsbereich noch nach oben hin ausgedehnt werden.

Nach Gleichung (22) wird das Korrekturglied Null für $\varphi = 0$ und $\varphi = 90°$. Bringt man also den Hallgenerator so in das Magnetfeld, dass grad B_z entweder in der x- oder y-Richtung liegt, so wird exakt die magnetische Induktion B_0 im Mittelpunkt $x = y = 0$ des Hallgenerators gemessen.

Im ungünstigsten Fall mit $\sin 2\varphi = 1$ nimmt das Zusatzglied in der geschweiften Klammer von (22)

seinen grössten Wert

$$\frac{\sigma R_h b^2}{24}|\text{grad}\, B_z|^2 \qquad (23)$$

an. Dieser ist für einen Hallgenerator aus InAs mit den bereits oben genannten Werten gleich $2\times 10^{-6}|\text{grad}\, B_z|^2$ Gauss, wenn grad B_z in Gauss cm^{-1} eingesetzt wird. Bei der Messung eines Magnetfeldes mit grad $B_z = 1000$ Gauss cm^{-1} kann daher der bei der Messung auftretende Fehler maximal 2 Gauss betragen. In gleicher Weise wie oben der Gültigkeitsbereich von (22) erweitert werden kann, lässt sich gemäss Gleichung (23) auch das Fehlerglied verkleinern.*

* Lässt man bei der Messung des Magnetfeldes eine Drehung des Hallgenerators um die z-Achse (vgl. Abb. 2) zu, so durchläuft die Hallspannung nach Gleichung (22) ein Maximum und ein Minimum, deren Mittelwert der magnetischen Induktion B_0 an der Stelle $x = y = 0$ und deren Differenz dem Quadrat des Gradienten von B_z proportional ist. Bei grossem Gradienten von B_z kann auf diese Weise B_0 exakt gemessen werden; ausserdem lässt sich |grad B_z| bestimmen.

Während für ein Magnetfeld mit konstantem Gradienten von B_z die Grundgleichungen exakt gelöst und die Hallspannung explizit angegeben werden kann, ist dies für ein beliebiges Magnetfeld nicht möglich. Wählt man aber zur Messung eines beliebigen, inhomogenen Magnetfeldes die Grösse des verwendeten Hallgenerators so, dass die Abweichung der Feldfläche $B_z(x, y)$ von der Tangentialebene im Punkt $x = y = 0$, $\bar{B}_z(x, y) = B_0 + (\partial B_z/\partial x)_0 x + (\partial B_z/\partial y)_0 y$ auf der gesamten Fläche des Hallgenerators klein gegen $\bar{B}_z(x, y)$ ist, dann gelten auch für die Messung beliebiger, inhomogener Felder die oben diskutierten Ergebnisse.

LITERATUR

1. W. Hartel, *Siemens-Z.* **28**, 376 (1954).
2. F. Kuhrt, *Siemens-Z.* **28**, 370 (1954).
3. H. J. Lippmann und F. Kuhrt, *Z. Naturf.* **13a**, 474 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 176–182. Printed in Great Britain

THE SCANSOR, A NEW MULTI-APERTURE RECTANGULAR-LOOP FERRITE DEVICE

S. DUINKER* and B. VAN OMMEN†

(*Received* 15 *January* 1960; *in revised form* 7 *March* 1960)

Abstract—The scansor consists of a multi-aperture plate of rectangular-loop ferrite material which is provided with a large number (e.g. 10–20) of separate output windings across which consecutive output pulses can be developed by driving the plate from one remanence position into the other by triangularly shaped pulses. The various output pulses are of rather uniform height but mutually delayed. Depending on the geometry of the scansor and on the slope of the driving current, the delay between pulses corresponding to any two adjacent single-turn output windings can be varied from 0·05 to 2μ with corresponding pulse heights of 10 to 0·2 V, respectively. Experiments are described showing that the response of certain output windings can be either suppressed or shifted in time by applying appropriate additional pulses. A few possible applications of scansors such as rapid scanning devices and code converters are briefly discussed.

Résumé—Le "scansor" consiste en une plaque d'une ferrite qui se caractérise par une courbe de hystérèse rectangulaire. La plaque est pénétrée de plusieurs ouvertures, dont chacune est liée par une bobine de débit. Des impulsions consécutives s'engendrent dans les bobines quand la plaque est commutée d'un extrême de rémanence à l'autre au moyen d'une impulsion triangulaire. Ces impulsions de débit sont d'une amplitude plutôt uniforme, mais elles sont mutuellement retardées.

Le retard entre les impulsions dans des bobines à unique tour adjacentes peut varier de 0·05 à 2·0 μsec (avec une amplitude correspondante de 10 à 0·2 V respectivement) selon la géométrie du scansor et l'inclinaison des impulsions triangulaires. Nous décrivons des expériences qui montrent que le débit de certaines bobines peut se faire ou supprimé ou deplacé temporellement par l'application d'additionelles impulsions appropriées. Nous délibérons brèvement quelques applications du scansor, telles que les déchiffreurs et les explorateurs rapides.

Zusammenfassung—Der "Scansor" besteht aus einer Ferrit-Platte mit Rechteckschleife; sie hat viele Öffnungen und ist mit einer grossen Anzahl (z.B. 10–20) getrennter Ausgangswicklungen versehen. An diesen Ausgangswicklungen kann eine Folge von Ausgangsimpulsen erzeugt werden, indem die Platte von einer Remanenzposition in die andere mittels Dreiecksimpulsen gesteuert wird. Die verschiedenen Ausgangsimpulse sind von ziemlich gleichmässiger Höhe, jedoch gegeneinander verzögert. Abhängig von der Geometrie des "Scansors" und dem Anstieg des Steuerstromes, kann die Verzögerung zwischen Impulsen, die zu jeweils zwei benachbarten Einzel-Ausgangswicklungen gehören, von 0·05 bis zu 2 μsec mit den zugehörigen Impulshöhen von 10 bis zu 0·2 V variiert werden. Experimente werden beschrieben, die zeigen, dass das Ansprechen der verschiedenen Ausgangswicklungen durch geeignete zusätzliche Impulse entweder unterdrückt oder zeitlich verschoben werden kann. Einige mögliche Anwendungen des "Scansors", wie z.B. schnelle Abtastgeräte und Kode-Umsetzer, werden kurz diskutiert.

1. INTRODUCTION

DURING the investigation into the possibilities of flux-canalization in multi-aperture plates of rectangular-loop ferrite, it was observed that, by applying triangularly shaped magnetizing-current pulses through one of the holes, useful output voltages could be obtained at output windings coupled to the various legs between the other holes or to other portions of the plate. These output voltages were of a reasonable shape and of a rather uniform height. Furthermore, the time-lag occurring between output pulses corresponding to portions

* Philips Research Laboratories, N.V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands.

† Now with N.V. Hollandse Signaalapparaten, Hengelo, Netherlands.

of the material at different distances from the driven aperture could be varied over a wide range by varying the slope of the driving-current pulses. These properties looked rather promising for application in various devices of the computer and switching technique.

In appearance the device has some resemblance with elements described in the literature under various names, such as: transfluxors[1], multi-aperture devices (MAD)[2], multi-path logical structures[3], etc. In property and in behaviour, however, it is in certain respects essentially different from this latter family of devices in which use is made of the fact that certain regions of remanent magnetization of mutually opposed polarity can be created in bodies consisting of ferrite material with a rectangular hysteresis loop, by applying magnetizing pulses of the appropriate polarity and preferably of a rectangular shape through the various apertures in the structure. This pattern of remanent magnetization, which corresponds to some configuration of blocked and unblocked flux paths, then serves as a basis for certain logical operations of the device by means of another set of electrical signals.

In the device to be discussed in the present paper it is the aim to derive at the various terminals output pulses which are mutually shifted in time by applying magnetizing pulses which are essentially of a non-rectangular shape and hence cause a propagating flux-reversing zone. The idea of the controlled speed of propagation of a flux-reversing zone is thought to be novel. For brevity and on account of its characteristic properties we have introduced the name "scansor"* for this device.

A number of experiments with scansors of ferroxcube VI E material (i.e. a rectangular-loop ferrite) will be discussed in order to explain the properties of the device. A few possible applications of scansors in various fields will be briefly indicated in conclusion.

2. PRINCIPLE OF OPERATION

The shape of a ferrite plate used in our first experiments is shown in Fig. 1(a). The principal dimensions are $21{\cdot}5 \times 12$ mm^2. The apertures were drilled ultrasonically and have diameters of 2 and

* Derived from the latin verb *scandere* = to climb, to scan.

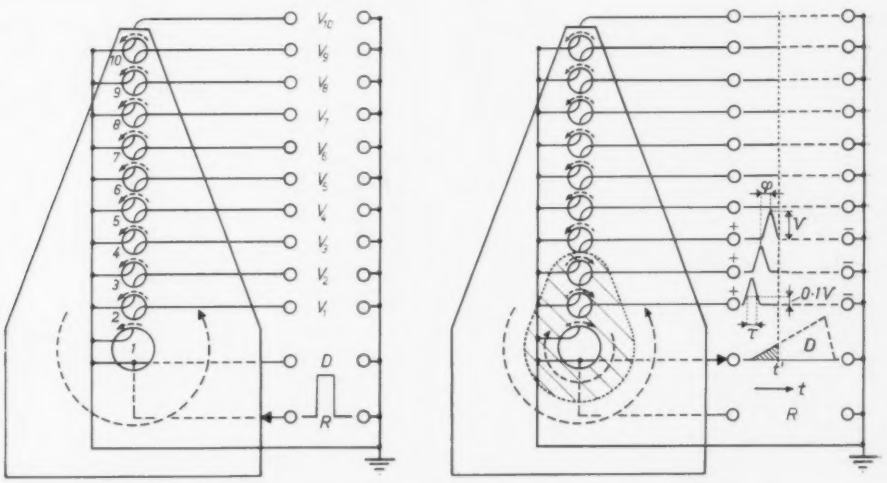


FIG. 1. Scansor with ten apertures provided with a reset winding R, a drive winding D and output windings $V_1, \ldots, V_{10}$: (a) Reset position in which the scansor is saturated by a rectangular reset pulse in a counter-clockwise direction with respect to the first aperture. (b) Drive position in which the magnetization in the shaded region around the first three apertures is reversed as a result of a triangular drive pulse D up to the time $t = t'$. The corresponding output pulses (height V, duration τ and mutual delay φ) are schematically indicated as a function of time.

1 mm for the large hole (1) and the smaller ones (2, ..., 10), respectively. The legs between the holes vary from 0·51 to 0·54 mm except for the first leg between holes 1 and 2 (0·58 mm) and the last leg between hole 10 and the edge of the plate (0·61 mm). The shortest distance from hole 1 to the edges amounts to 4·95 mm. The thickness of the plate is 0·9 mm.

Two windings, a "drive" winding D and a "reset" winding R, are threaded through the first aperture. Single-turn output windings are wrapped around each leg between adjacent apertures (1–2, 2–3, ...).

Through the reset winding a rectangular current pulse is sent, of a strength which is sufficient to saturate the material, e.g. in a counter-clockwise direction. After this pulse the remanent magnetization in all the legs at the upper side of the first aperture is directed to the left as indicated in Fig. 1(a). If afterwards a pulse, of a polarity opposite to that of the reset pulse and of a triangular shape, is sent through the drive winding, then, as soon as the critical field-strength of the material at the circumference of the first aperture is surpassed, a circular region of a reversed magnetization is formed, the boundary of which will travel outwards at a speed which is dictated by the slope of the driving current. Each time that the boundary passes a leg, the magnetization in that leg is reversed, thereby causing an output pulse to occur across the corresponding winding. Such a situation, where the boundary has just passed the third leg, is pictured in Fig. 1(b) in which also the responses of the first three output windings are schematically indicated as a function of time. Eventually, the magnetization in all the legs has consecutively been reversed, leaving the material saturated in a clockwise direction. A further increase in driving current will then not have any further effect.

For a proper functioning of the scansor the smallest cross-sectional area of the plate between the first aperture and one of the edges should be equal to or greater than the sum of the cross-sections of the legs, otherwise it is impossible to reverse the magnetization in those extreme legs (8, 9, ...) for which there is no longer a closed flux path in the material around the first aperture in which the magnetization can also be reversed. The pointed shape of the plate was chosen in order to constrain the flux of the outer legs to the shortest path.

3. EXPERIMENTAL RESULTS

We shall first discuss some experimental results obtained with the scansor specified in the previous section. Afterwards we shall present some data pertaining to a scansor with a greater number of apertures.

The quantities which are of interest for a discussion of the behaviour of the scansor are indicated in Fig. 1(b) as follows: the maximum value V of the output pulses, the time delay φ between pulses corresponding to adjacent legs, and the switching time τ, i.e. the duration of a pulse measured at the 10 per cent level.

In the series of photographs represented in Fig. 2(a), the voltage across the single-turn output winding of each leg is shown. These pictures correspond to a drive current D with a rising slope of 1 A/μsec as is reproduced in the lowest two photographs. The two other photographic series will be dealt with in the next section.

We see that the height of the output pulses attains a maximum value at the sixth leg and then falls rather quickly so that no response occurs at the tenth leg. This is due to the fact that in the geometrical shape of the plate shown in Fig. 1 the requirement for a closed flux path of uniform width is not satisfied for the outer limbs so that these are only partially switched, if at all. This effect is not observed in the scansor that will be discussed later on.

The delay φ between consecutive output pulses, which is seen to be approximately 0·1 μsec, decreases somewhat for legs situated at a greater distance from the first aperture. Also, the switching time τ is not constant for the various legs. To some extent these effects are due to the geometrical shortcomings of the plate.

In order to obtain a quantitative insight into the dependence of V, φ and τ on the slope S of the drive current, a series of measurements have been carried out for the responses corresponding to the second leg. These data are collected in Fig. 3 on a log–log scale. We see that the logarithm of the output voltage V is linearly related to the logarithm of the current slope S. The same holds true for the switching time τ and the delay φ. The following relations are derived from the figures: $V = 1{\cdot}2S^{+0{\cdot}61}$ V, $\tau = 0{\cdot}33S^{-0{\cdot}56}$ μsec, $\varphi = 0{\cdot}14S^{-0{\cdot}70}$ μsec, where S denotes the value of the slope of the rising portion of the drive-current pulse in A/μsec. Of course, the numerical values in these expressions

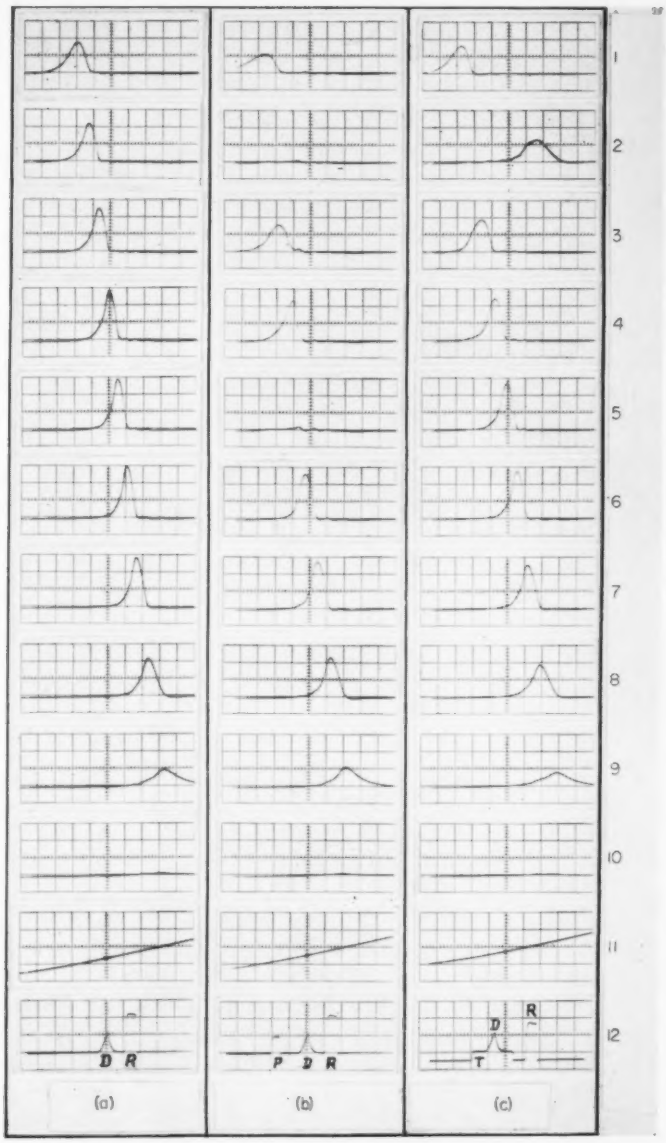


FIG. 2. Oscillograms showing actual pulses corresponding to the scansor schematically represented in Fig. 1 for a drive current with a slope of 1A/μsec and under three different conditions: (a) normal scansor response; (b) suppression of the response of the second and fifth leg by a preset pulse P; (c) time-shift of the response of the second leg by a shift pulse T. The oscillograms 1–10 show the output pulses across the individual single-turn output windings (vertical scale: 1 unit = 0·5 V; horizontal scale: 1 unit = 0·2 μsec), oscillograms 11 represent the slope of the drive current (vertical scale: 1 unit = 1 A; horizontal scale: 1 unit = 0·2 μsec) and oscillograms 12 represent the total energizing pattern showing reset pulse R, drive pulse D, preset pulse P and time-shift pulse T; actually the polarities of R and T are opposite to those of D and P (vertical scale: 1 unit = 2 A; horizontal scale: 1 unit = 5 μsec).

[*facing p.* 178

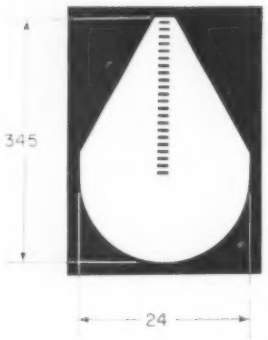


FIG. 4. Diagram of a scansor with twenty-two rectangular apertures.

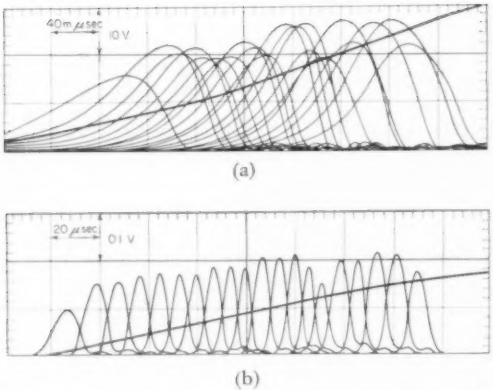


FIG. 5. Voltages across five-turn output windings of the first twenty legs of the scansor represented in Fig. 4, together with the slope S of the drive current: (a) for $S = 15$ A/μsec; (b) for $S = 0{\cdot}004$ A/μsec.

depend largely upon the geometrical shape of the plate. For instance, higher output voltages and lower values for τ and φ are obtained if, respectively, the plate is made thicker and the pitch of the apertures is reduced.

For comparison a number of measurements were carried out with a scansor of the same general shape as the one represented in Fig. 1, but with a larger number of apertures (22) and for which the requirement imposed on the cross-sectional areas in the various parts of the plate, in order to secure a closed flux path for all the legs, was indeed satisfied (see Fig. 4). Unfortunately, the widths of the individual limbs between the various apertures could not be kept sufficiently constant during the ultrasonic drilling operation.

With a 20-positions switch, the output voltages of the first twenty legs across five-turn windings can be displayed in a single oscillogram. Redrawings of such oscillograms for two different values of the slope S of the drive current are shown in Fig. 5(a) and (b), viz. for $S = 15\text{A}/\mu\text{sec}$ and $S = 0\ 004\text{A}/\mu\text{sec}$ respectively. It is seen that, after correcting the voltage amplitudes for the variations in the width of the legs and for the departure of the current slope from linearity, output pulses of a fairly uniform shape and height are obtained. From these results it seems justifiable to deduce that the number of apertures, and hence the number of outputs, can be still further increased, although this was not experimentally verified.

4. SUPPRESSION AND TIME-SHIFT OF LEG RESPONSE

In the experiments described thus far, the resetting and the subsequent driving conditions are, in principle, the same for the various individual legs. For various applications it would be of interest to know whether responses of certain legs can be suppressed or time-shifted by counteracting or inhibiting fields which are introduced before or during the driving pulse. To get an impression of these possibilities, two experiments, which are illustrated by the photographic series shown in Fig. 2(b) and (c), respectively, have been carried out, using again the scansor represented in Fig. 1. For a better comparison, the same slope of the driving current for both cases was chosen as in the experiment pictured in Fig. 2(a), i.e. 1 A/μsec.

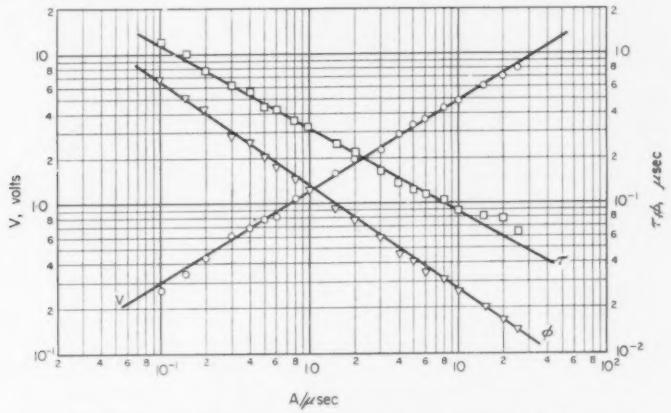


FIG. 3. Dependence of the output voltage V, the delay φ and the switching time τ on the slope S of the drive current.

Case (i)—*Pre-switching of legs*

In addition to the reset winding and the drive winding through the first aperture, and the single-turn output windings coupled to the various legs, a "pre-switching" winding is linked with the second and the fifth legs only, these legs being arbitrarily chosen. Through this winding a current pulse P is sent which comes after the reset pulse R but precedes the drive pulse D. The shape of these is shown in the last photograph in Fig. 2(b); it should be borne in mind, however, that actually the polarity of the R-pulse is opposite to that of the pulses P and D.

The *P*-pulse causes a reversal of the magnetization of the second and fifth legs and of the corresponding closed flux paths surrounding the first aperture. The output voltages accompanying the flux reversals in these legs are not shown in the photograph.

Next, when applying the triangularly shaped *D*-pulse, the zone of reversing magnetization, in travelling from the first hole to the last hole, passes the regions corresponding to the second and fifth leg in which the magnetization is already reversed. Hence, output voltages are consecutively obtained from the various legs except from the second and the fifth leg. This is shown in the photographic series represented in Fig. 2(b).

In comparing the photographic series of Fig. 2(a) and (b) we see that the pre-switching has some effect on the output voltages and the switching times of adjacent legs. However, the discrimination ratio between the output voltages of selected (i.e. pre-switched) legs and non-selected legs is still very high. If the pre-switching is extended to more than two legs, the effect on the shape of the pulses corresponding to non-selected legs will increase, but, even in the case that all the legs but one are pre-switched, a useful output results for the non-selected leg.

Case (ii)—*Time-shift of leg response*

In addition to the reset winding and the drive winding through the first aperture and the single-turn output windings coupled to the various legs, a "time-shifting" winding is linked to the second leg, this leg being again arbitrarily chosen. In the last photograph of Fig. 2(c) the shapes of the pulses through these windings are shown together; actually, the polarity of the time-shifting pulse *T* and of the reset pulse *R* is the same, and opposite to that of the drive pulse *D*. The duration of the *T*-pulse was chosen larger by a factor of approximately 3 than that of the *D*-pulse in order to ensure that the time-shifting field is present in the second leg all the time that the plate is driven.

The operation is now as follows. Since the magnetization of the second leg is kept in the reset position by the *T*-pulse during the interval of time that the *D*-pulse is effective, the zone of reversing magnetization comes to a stop at the second leg so that no pulse results at the second output winding. Beyond the second leg a new zone is created which travels further outwards, thereby again creating voltages across the corresponding output windings, as is shown in Fig. 2(c). However, as the driving current is increasing, the field strength in the second leg decreases, since the polarities of the pulses *T* and *D* are opposite. The field strength goes through zero and changes sign as soon as the *D*-pulse starts to over-rule the *T*-pulse. Eventually, the blocking action of the *T*-pulse is completely cancelled and, since the *D*-pulse is still in its rising portion, the critical field strength of the second leg will be surpassed and the magnetization in it will be reversed causing a pulse across the second output winding. For the current conditions chosen in the experiment the reversal of the second leg is approximately coincident with that of the eighth leg, corresponding to a delay of about 0·8 μsec. By variation of the magnitude of the *T*-pulse and/or the slope of the *D*-pulse even larger delays can be obtained and, of course, also smaller delays.

In this experiment there is also a small effect on the response of adjacent legs, as can be deduced from a comparison of the photographic series represented in Fig. 2(a) and (c), but this effect is smaller than in the preceding experiment. It will be clear that the delaying procedure need not be confined to a single leg.

The advancing of pulses is also possible by reversing the polarity of the *T*-pulse, so that it produces an aiding field instead of a counteracting field for the driving field. If the magnitude of this aiding field is kept small enough so that the material of the corresponding leg is kept in the reset condition, the critical field strength will be reached sooner than would have been the case without this field. If the aiding field is made too strong so that the leg switches, the experiment described in case (i) is obtained.

5. POSSIBLE APPLICATIONS

On account of the experimental results described in the foregoing, numerous applications of scansors in different fields suggest themselves immediately. We shall briefly indicate a few of these possibilities.

An obvious application is to use the scansor as a scanning device that distributes output pulses consecutively to an array of inputs, an operation which can be useful in various fields of switching technique. For instance, an electronic version of a

multi-output mechanical rotary switch of the type which is frequently used in telephony can be made by connecting each output of the scansor directly to a corresponding input of the system that has to be scanned. If this connexion is made via a logical element (e.g. an and-gate) and if provision is made for marking potentials to be effective at each of the inputs, then the switch can be made to stop at a particular input for which the marking potential exceeds a certain threshold value by deriving from the and-gates a signal that abruptly interrupts the driving-current pulse of the plate. There are various advantages of such an all-electronic scanning system over the conventional mechanical switches, namely, absence of moving parts, longer life, possibility of much higher speeds, etc.

Another field where the scansor might present a solution is that of the scanning of electro-luminescent display panels. Various proposals have been made[4] to employ the effect of electro-luminescence for the display or for the temporary storage of visual patterns (e.g. for television, radar, data handling) by arranging individual electro-luminescent cells as picture elements in a matrix form between cross-bar electrodes. A particular picture element is selected by energizing the electrodes crossing at that element. For the continental European 625-line television system, about 600 picture elements of each line have to be scanned in 64 μsec, corresponding to a delay of 0·1 μsec between consecutive picture elements. This rapid scanning procedure could be obtained with a relatively small number of inexpensive scansors, each provided with multi-turn output windings across which peakvoltages up to 100 V can be developed under loading conditions such as are encountered for a display panel and which are of the order of 100 pF in parallel with several hundreds of ohms. Alternative methods, using electron tubes, transistors, etc., would be prohibitively expensive, whereas the various types of delay line would not generally satisfy the severe requirement of unattenuated uniform output pulses for such a great number of taps.

An entirely different application is the use of the scansor as an analogue-to-digital converter. Square-wave signal of equal height but of different duration will produce, after integration, linearly increasing signals with an equal slope for those different intervals of time. If these latter signals are used for driving the scansor from the reset position, then the number of outputs of the scansor that will deliver a pulse (or, if the outputs are all connected in series, the number of pulses appearing at the common output winding) is related to the duration of the original pulse from which the driving current pulse was derived.

The scansor can also be used for the generation of coded signals, either by connecting the output windings in series according to the desired pattern (e.g. series-connection of the first, second and fifth output winding for the generation of the pattern 11001), or by connecting all the output windings in series and suppressing the response of the legs corresponding to a zero in the way described in case (i) of Section 4. Another possible application in this field is the conversion of code patterns, e.g. secrecy codes.

The time-shift of output pulses as described in case (ii) of Section 4 offers possibilities for applications such as pulse-code modulation, but also for converting a d.c. signal into a pulse signal for which the time-separation of a pair of successive pulses between the reset pulses is related to the strength of the d.c. signal.

The few examples which were just briefly discussed and which are by no means exhaustive will suffice to give an impression of the applicability of scansors in various electronic fields. The number of possibilities can be further increased by considering scansors under slightly different conditions (e.g. a drive current with a non-linearly rising slope, or a slope rising according to a multi-step function) or by using scansors of a somewhat different shape (e.g. with a non-uniform pitch for the consecutive apertures, with apertures of another shape, or with apertures arranged in more than one row). Such modifications will not be dealt with further in this paper although a few of them have produced interesting experimental results.

6. CONCLUSION

The experiments with two prototypes of scansor described in the foregoing sections indicate that it is possible to build scansors with more than 25 apertures and with a delay between the output pulses at adjacent output windings in the interval of 0·1–1 μsec and even beyond this range. By choosing thicker plates and multi-turn output

windings, output pulses with peak values of 10–00 V and more may be obtained.

If the scansor plates, instead of being ultrasonically drilled, are pressed and sintered to shape, the device presents a rather inexpensive alternative to various problems in the switching and the data-handling technique, such as the rapid scanning of multi-input systems, the analogue-to-digital conversion, and the generation or conversion of code patterns.

REFERENCES

1. J. A. RAJCHMAN and A. W. LO, *R.C.A. Rev.* **16,** 303, (1955); *Proc. I.R.E.* **44,** 321 (1956).
2. H. D. CRANE, *Proc. I.R.E.* **47,** 63 (1959).
3. S. A. ABBAS and D. L. CRITCHLOW, *National Convention of the Institute of Radio Engineers*, Part 4, p. 263 (1958); N. F. LOCKHART, *Ibid.* Part 4, p. 268 (1958).
4. J. A. RAJCHMAN, G. R. BRIGGS and A. W. LO, *Proc. I.R.E.* **46,** 1808 (1958); E. A. SACK, *Ibid.* **46,** 1694 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 183–187. Printed in Great Britain

GERMANIUM p–n-JUNCTION–TUNNEL-JUNCTION COMBINATION DEVICES

I. A. LESK and H. A. JENSEN

General Electric Company, Semiconductor Products Department, Syracuse, New York

(*Received* 26 *February* 1960)

Abstract—Semiconductor devices employing both normal p–n junctions and tunnel junctions (designated p_t–n) can utilize properties of a tunnel junction to impart new or improved characteristics to the overall structure. A p_t–n junction, in the low-voltage region preceding the negative resistance, has current carried across the junction barrier primarily by electron tunneling, with very little minority-carrier injection. In the higher-current region following the negative resistance, minority-carrier injection occurs. Germanium p_t–n–p transistors, having a tunnel-junction emitter and a normal-junction collector, were made by a combination of diffusion and alloying processes. They exhibit a current gain of essentially zero at low emitter currents, jumping abruptly to a finite value at the tunnel-junction peak current. This effect was applied to control the characteristics of four region switches, for which p_t–n–p–n structures were used. The p_t–n–p–n triode can be made to have its switching current equal to the tunnel-junction peak current, and its hold current equal to the tunnel-junction valley current. A base current equal to the peak current may be used to turn the devices on. These structures can be made to have quite stable properties from −70°C to +100°C, since germanium tunnel diodes show little change in characteristics over this temperature range.

Résumé—Des dispositifs employant aussi bien des jonctions normales p–n que des jonctions tunnel peuvent mettre en oeuvre des propriétés d'une jonction tunnel pour donner des caractéristiques nouvelles ou améliorées à la structure de l'ensemble. Une jonction p_t–n, dans la région du bas voltage précédent la résistance négative, a un courant qui passe à travers la barrière de la jonction grace à l'effet tunnel des électrons avec une injection de très peu de porteurs minoritaires. Dans la région du courant plus élevé au-delà de la résistance négative, il apparait une injection de porteurs minoritaires. Des transistors p_t–n–p au Germanium ayant une jonction tunnel comme émetteur et une jonction normale comme collecteur ont été faits en combinant les processus de diffusion et d'alliage. Ils présentent un gain en courant à peu près nul à bas courant émetteur et qui saute brutalement à une valeur finie pour le courant du pic de la jonction tunnel. Cet effet fut utilisé à controler les caractéristiques de commutateurs à quatre régions pour lesquels étaient utilisées des structures p_t–n–p–n. La triode p_t–n–p–n peut avoir son courant de commutation égal au courant de pic de la jonction tunnel, et son courant d'arret égal au courant minimum de la jonction tunnel. Un courant de base égal au courant de pic peut être utilisé pour amorcer le dispositif. Ces structures peuvent avoir des propriétés stables entre −70°C et +100°C puisque les diodes tunnel en Germanium ont leurs caractéristiques très stables dans cette région de température.

Zusammenfassung—In Halbleiterbauelementen, die sowohl normale p–n-Übergänge als auch Tunnelübergänge (als p_t–n-Übergänge bezeichnet) enthalten, können die Eigenschaften eines Tunnelüberganges ausgenutzt werden, um dem Bauelement neue oder verbesserte Charakteristiken zu verleihen. Der Strom über einen p_t–n-Übergang ist bei kleinen Spannungen, die unterhalb des Bereiches des negativen Widerstandes liegen, hauptsächlich ein Tunnelstrom. Die Injektion von Minoritätsträgern ist in diesem Bereich sehr klein. In dem Bereich höherer Stromstärke, der auf den negativen Widerstand folgt, tritt Minoritätsträger-Injektion auf. Germanium-p_t–n–p-Transistoren, die einen Tunnelübergang als Emitter und einen Normalübergang als Kollektor haben, wurden durch eine Kombination von Diffusion und Legierungsprozess hergestellt. Sie zeigen eine Stromverstärkung, die bei kleinen Emitterströmen praktisch Null ist und die abrupt auf einen endlichen Wert springt, sobald man das Strommaximum des Tunnelüberganges erreicht. Dieser Effekt wurde angewendet, um die Charakteristiken von Vierschichtkristallen zu beeinflussen; es wurden dafür p_t–n–p–n-Strukturen benutzt. Die p_t–n–p–n-Triode lässt sich so herstellen, dass ihr Schaltstrom gleich dem Strommaximum und ihr Haltestrom gleich dem Stromminimum des

Tunnelüberganges ist. Durch einen Basisstrom, der gleich dem Strommaximum ist, kann die Triode eingeschaltet werden. Diese Strukturen können so gemacht werden, dass ihre Eigenschaften zwischen −70°C und +100°C sehr stabil sind, da die Kennlinien der Germanium-Tunneldioden sich in diesem Temperaturbereich nur wenig ändern.

INTRODUCTION

THE tunnel diode[1–5] is an abrupt p–n junction with very large (degenerate) impurity concentrations in both n-type and p-type regions. When "forward" biased, it displays a negative resistance region, as shown in Fig. 1. If a tunnel diode is driven from a high-impedance source, its apparent characteristic jumps at peak and valley points, as indicated in Fig. 2. Three regions in a tunnel-diode characteristic are shown in Fig. 3. In the tunnel region, current is carried across the barrier primarily by quantum-mechanical tunneling of electrons. In the injection region, the tunnel diode is behaving much like a normal p–n-junction diode, and minority carriers are injected across the barrier. Hole and electron concentrations, and their respective currents, are determined by impurity concentrations, diffusion lengths and the device geometry. The excess-current region, in which the tunnel and injection regions merge, has current carried by both mechanisms. Hence, when the tunnel diode is driven from a high-impedance source, the jump to a higher voltage at the peak current involves a sudden change in the nature of the current, from current carried almost entirely by electron tunneling to one in which the degree of injection is large. The jump to a smaller voltage at the valley involves a sudden change from an at least partially injecting mode to one in which the current is carried preponderantly by means of electron tunneling.

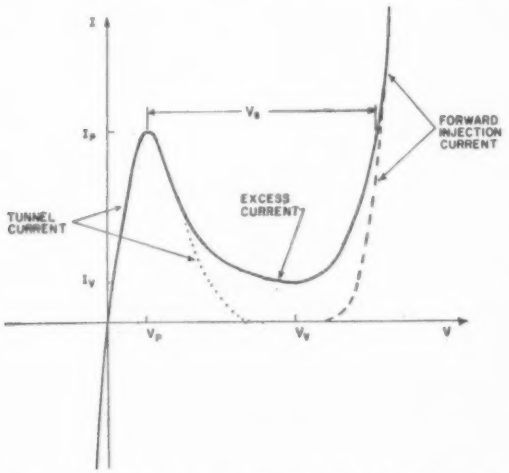


FIG. 3. Three regions in a tunnel-diode characteristic.

When a tunnel diode is placed in series with one or more of another type of p–n-junction device, the load presented to the tunnel diode is generally of an impedance high enough to cause switching, because the tunnel diode has such a low impedance (Fig. 2). Hence, in a structure containing in series a tunnel junction and one or more other p–n junctions, voltage across the tunnel junction will suddenly switch, and current through the tunnel junction will suddenly change character, as current through the device is increased past the peak, or decreased below the valley, of the tunnel junction. These properties are utilized in the devices that follow.

For convenience, a tunnel p–n junction will be designated p_t–n.

The p_t–n–p transistor

A p_t–n–p transistor (p–n–p transistor with a tunnel-junction emitter) may be made as shown in Fig. 4. Into a p-type germanium wafer was diffused an n-type skin, the outer part of which was degenerate. A p-type emitter alloy region, also degenerate, was made in the n-type skin. The p-type emitter may either remain within the degenerate n-type region, Fig. 4(a), or penetrate to the non-degenerate region, Fig. 4(b). A base contact was made by means of an n-type alloy to the diffused region, and the collector was a p-type region to the parent wafer.

The emitter-base characteristic for either structure of Fig. 4 is that of a tunnel junction, most of the emitter current flowing in a thin layer adjacent to the surface. If part of the n-type diffused layer is etched away, the impurity concentration on the base side of the emitter-base p–n junction at the surface (the most highly conducting path) is reduced. Hence, the emitter-base tunnel characteristic may be controlled by etching, as well as by the initial diffused impurity profile and the emitter-

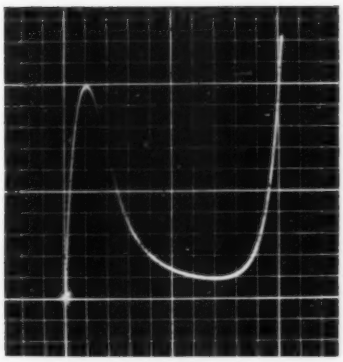


FIG. 1. Characteristics of a germanium tunnel diode, forward bias only. 0·1 mA/div. vertically, 0·05 V/div. horizontally.

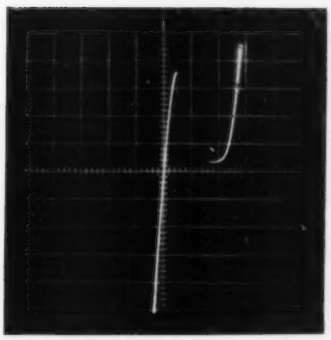


FIG. 2. Apparent characteristics of a tunnel diode when driven from a high-impedance source. Jumps occur at peak and valley currents. Origin is at centre. 0·5 mA/div. vertically, 0·2 V/div. horizontally.

[*facing p.* 184

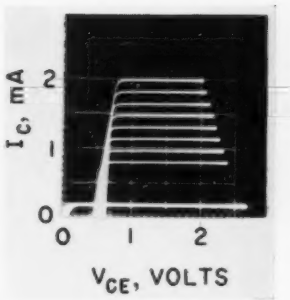


FIG. 5. Collector characteristic (common emitter) for a germanium p_E–n–p transistor having its emitter entirely within the degenerate base region. Bottom curve—I_B = 0, 50, 100, 150, 200 mA. Upper curves—I_B = 250–600 mA in 50-mA increments.

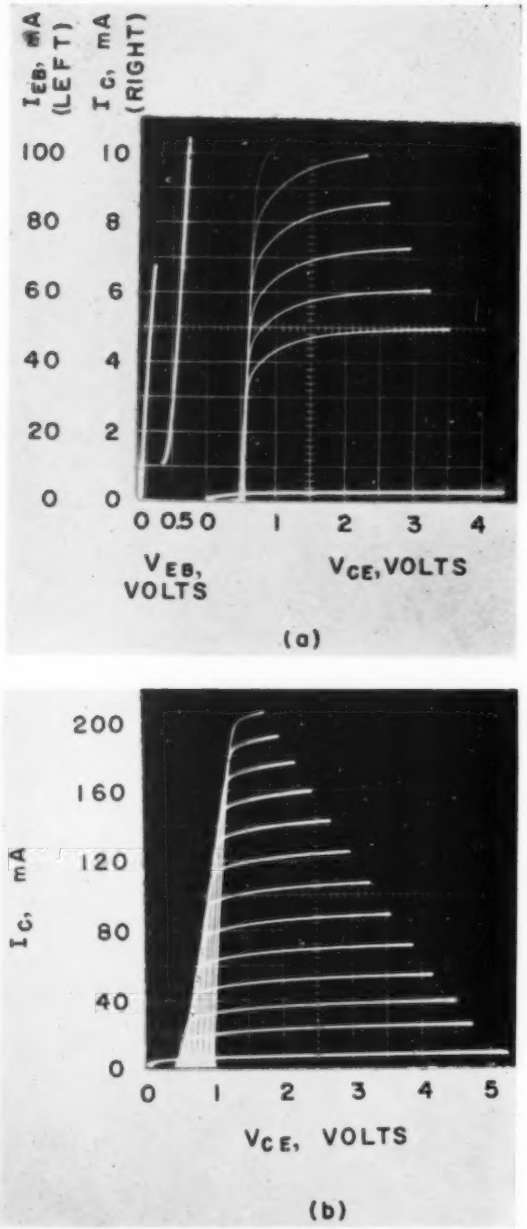


FIG. 6. Collector characteristic (common emitter) for a germanium p_t–n–p transistor having its emitter mostly within the non-degenerate base region. (a) Left: emitter-base tunnel characteristic. Right: bottom curve: $I_B = 0$, 10, 20, 30, 40, 50, 60 mA; upper curves: $I_B = 70$–120 mA in 10-mA steps. (b) Bottom curve: $I_B = 0$, 50 mA; upper curves: $I_B = 100$–600 mA in 50-mA steps.

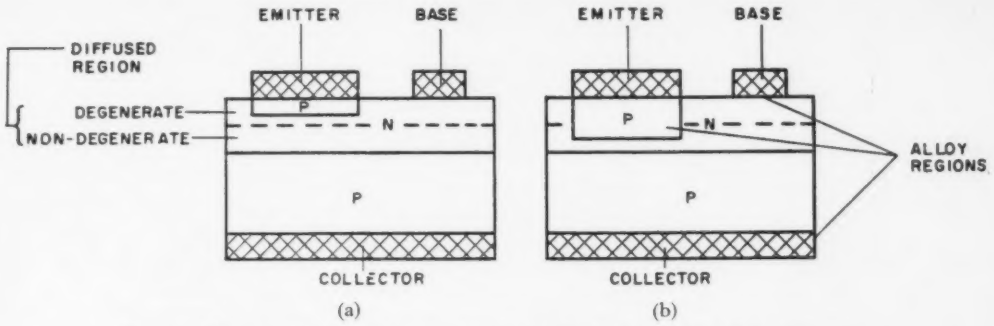


FIG. 4. Two types of p_t–n–p transistors. (a) Emitter entirely within degenerate base layer. (b) Emitter mostly within non-degenerate base layer.

alloy-region diameter. The base-collector junction is a normal p–n junction because of the small impurity gradient there.

Consider the transistor of Fig. 4 connected in the common-emitter configuration. If some base current is drawn, the emitter is slightly forward biased, and its current is all carried by electrons tunneling from emitter to base. Injection efficiency, and hence current gain, is extremely low. When the emitter current reaches the tunnel peak, the emitter voltage will jump and the emitter suddenly start to inject holes into the base. The base current required for this is also equal to the tunnel-junction peak current because of the very low alpha. When the emitter switches to the injecting mode, its current switches paths; instead of flowing in a thin surface layer at the emitter periphery, it will concentrate at the bottom of the emitter junction. Injection efficiency is now finite, and the collector current will jump by a value βI_B, where β is the d.c. common-emitter current gain for the emitter in the injecting mode, and I_B is the total base current.

Fig. 5 shows the collector characteristics (common emitter) for a device having a geometry like that shown in Fig. 4(a). The scale is 0·5 mA/div., 0·5 V/div. The lower curve is I_{CEO}, and I_C for $I_B = 50$, 100, 150 and 200 mA. The bottom curve of the group is for $I_B = 250$ mA, and the increment of base current is 50 mA for each higher curve. The current gain is zero, and jumps to 0·004 at the emitter-tunnel peak current (between 200 and 250 mA). The voltage swing of the emitter junction in jumping from the tunnel to the injection region is shown as a displacement (about 0·5 V) between the origins of the bottom curve and the group of curves.

Fig. 6 shows the collector characteristics (common emitter) for a transistor having a structure like that shown in Fig. 4(b). The emitter-base characteristic, with a peak current of 68 mA, appears on the left, Fig. 6(a). On the right, Fig. 6(a), it may be seen that I_{CEO} is the only current flowing until the base (and hence emitter) current exceeds the peak current. The first trace of the group is for $I_B = 70$ mA. When triggered on, the alpha of this transistor is above 0·1. Fig. 6(b) shows the collector characteristics at higher current. The larger saturation voltages for higher-current curves are due to transverse base biasing caused by the extremely large base currents, each larger value of which gives a measurably larger collector floating potential. Note that the $I_B = 50$ mA trace coincides with the I_{CEO} curve. At high collector currents, the gain approaches 0·4.

The current gain of another germanium p_t–n–p transistor as a function of current, is plotted in Fig. 7. This curve was obtained using a d.c. drive so that the transistor could be held in any operating (common-emitter) condition. The value of α jumps suddenly at the emitter-tunnel peak current, and then increases linearly with I_E (in the low-current region). Upon decreasing I_E, α decreases linearly until I_V is approached, in which region a rapid drop occurs. The shape of this curve may indicate that some tunneling occurs in the injection region even at currents well past the p_t–n peak. However, the linear increase of β with I_E may also be interpreted as a minority-carrier lifetime increase with injection level.

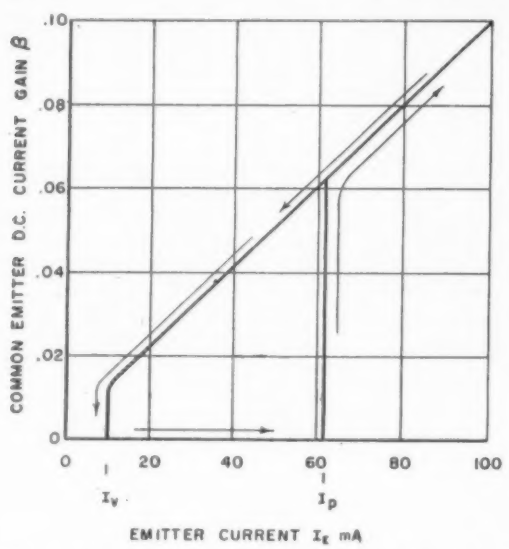


FIG. 7. Common emitter-current gain as a function of emitter current for a germanium p_t–n–p transistor. Jump occurs at emitter-tunnel peak current, drop at emitter valley current.

The p_t–n–p–n switch

The p_t–n–p–n switch is constructed much the same as the p_t–n–p transistor, except that the n-type diffused layer on the back of the wafer is not removed. Ohmic contact to this layer is made by means of an n-type alloy. Geometries corresponding to those in Fig. 4 may then be made as shown in Fig. 8. In Fig. 8(a) is illustrated a design in which the emitter is entirely within the degenerate base layer. Fig. 8(b) shows the case in which most of the emitter is within the non-degenerate base layer.

The p–n–p–n switch relies for its operation upon the fact that its current gain may be made to be low at small currents, and increase with current until it reaches unity (at the switching current) without the help of avalanche multiplication. The current gain of a p–n–p–n switch is equal to the sum of the current gains of its separate p–n–p and n–p–n transistor sections, each sharing a common collector junction. Several physical mechanisms have been shown to result in a rather slow increase of transistor current gain with current,[6–10] and more recently a sudden shift of current path in a shorted emitter structure has been utilized.[11]

In the p_t–n–p–n switch, the increase of current gain with current is also sudden and takes place when the emitter jumps to an injecting mode as its peak tunnel current is exceeded. If this increase is sufficient to make the total current gain greater than unity, the p_t–n–p–n device switches to its low-impedance state. Upon reducing the current, the switch will turn off when the total current gain drops to unity. This cannot be lower than the p_t–n-junction valley current, but may be higher, depending on other factors in the device design. When used as a triode, the p_t–n–p–n switch may be turned on by a gate current equal to the p_t–n junction peak current (switching current).

Fig. 9 shows the characteristics of a germanium p_t–n–p–n triode at three temperatures. At the left in each picture is shown the p_t–n junction characteristic to the same current scale (but to a greatly expanded voltage scale) as the p_t–n–p–n trace. In

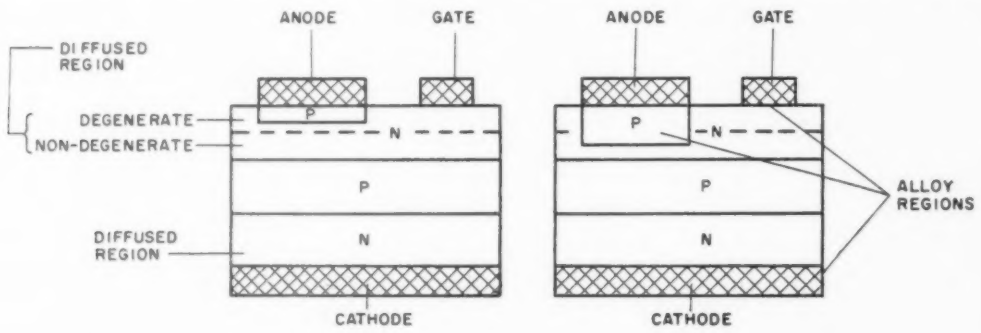
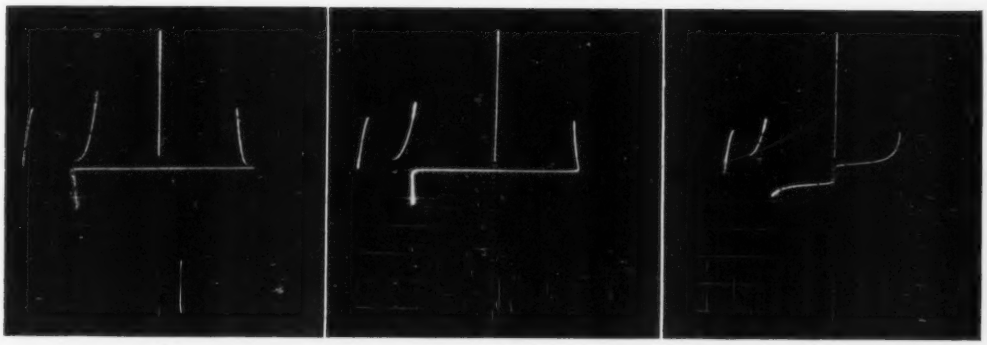


FIG. 8. Two types of p_t–n–p–n switches. (a) Anode entirely within degenerate region. (b) Anode mostly within non-degenerate region.



−70°C 27°C 100°C

FIG. 9. Germanium p_t–n–p–n switch characteristic, with tunnel-emitter characteristic at left. 10 mA/div. vertically. 10 V/div. horizontally—p_t–n–p–n trace. 0·2 V/div. horizontally—p_t–n trace.

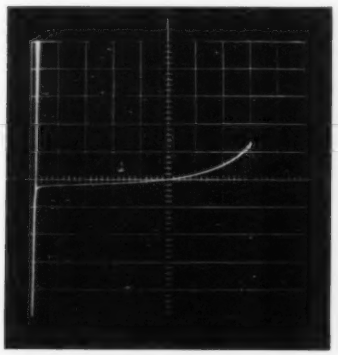
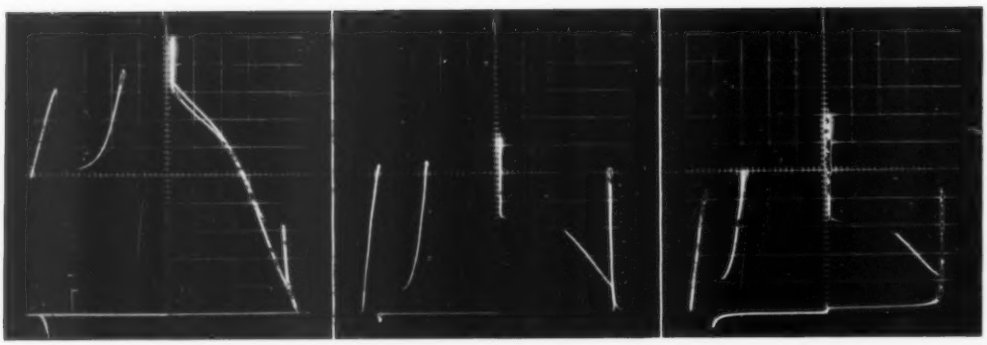


FIG. 10. Germanium p_t–n–p–n switch at 135°C. Hold current is less than leakage current, so negative-resistance region does not exist. 5 mA/div. vertically. 2 V/div. horizontally.

[*facing p.* 186



77°K 27°C 100°C

FIG. 11. Germanium p_t–n–p–n switch characteristics, with tunnel-emitter characteristic at left. $I_V < I_H$ (p–n–p–n) $< I_P$ at 27°C and 100°C. $I_V < I_P < I_H$ (p–n–p–n) at 77°K. 5 mA/div. vertically. 5 V/div. horizontally for p_t–n–p–n. 0·2 V/div. horizontally for p_t–n.

this case, control of the p_t–n junction over the device performance is strong. The p_t–n peak current determines the switching current over the entire temperature range. The p_t–n valley current determines the hold current at the upper two temperatures. At $-70°C$, the switch turns off at a current slightly higher than the valley current. This is probably because the minority-carrier lifetime is lower at this temperature and, together with the reduced injection efficiency as the valley current is approached, causes the current gain to drop to unity before the valley is reached (and switching to the tunnel region occurs).

If gate current is used to trigger this p_t–n–p–n device, the characteristic is practically unchanged for base currents less than the p_t–n-junction peak current, and switching to the low-impedance state occurs when the base current is equal to or greater than the p_t–n-junction peak current.

The ability of a p_t–n junction to keep the p_t–n–p–n switch from turning until its peak current is exceeded is illustrated more dramatically in Fig. 10. In this case, the device was heated to 135°C, at which temperature the leakage current was larger than the hold current. The switch still has a high-impedance region because the leakage current is less than the p_t–n peak current, and thermal runaway has been avoided.

In Fig. 11 are shown the characteristics for a p_t–n–p–n switch over which the tunnel emitter has less control than in Fig. 8. At room temperature, the p_t–n peak coincides with the switching current. As current is reduced, the device starts to turn off before the p_t–n valley is reached, and the characteristic follows a high-voltage curve along which avalanche multiplication contributes towards increasing the current gain to unity. Final turn-off is reached, however, at the valley current. At liquid-nitrogen temperature, left, avalanche multiplication is required to a current much higher than the p_t–n peak to keep the current gain at unity. For this reason the device will turn on as the p_t–n peak is exceeded only if the series resistance is so small that the current can jump to greater than about 45 mA.

CONCLUSIONS

The tunnel junction may be utilized along with other p–n junctions in the same structure to make complex devices, with novel, useful characteristics. The examples discussed in this paper utilize the change in character of the tunnel-junction current in going from the tunnel to the injection regions to control the characteristics of the p–n–p and p–n–p–n structures.

REFERENCES

1. L. Esaki, *Phys. Rev.* **109,** 603 (1958).
2. H. S. Sommers, Jr., *Proc. I.R.E.* **47,** 1201 (1959).
3. I. A. Lesk, N. Holonyak, Jr., U. S. Davidsohn and M. W. Aarons, *I.R.E. Wescon Convention Record,* Part 3, p. 9 (1959).
4. R. F. Rutz, I.B.M. *J. Res. Dev.* **3,** 372 (1959).
5. N. Holonyak, Jr., I. A. Lesk, R. N. Hall, J. J. Tiemann and H. Ehrenreich, *Phys. Rev. Letters* **3,** 167 (1959).
6. J. L. Moll, M. Tanenbaum, J. M. Goldey and N. Holonyak, Jr., *Proc. I.R.E.* **44,** 1174 (1956).
7. C. J. Sah, R. N. Noyce and W. Shockley, *Proc. I.R.E.* **45,** 1228 (1957).
8. C. W. Mueller and J. Hilibrand, *I.R.E. Transactions on Electron Devices,* ED-**5,** 3 (1958).
9. R. W. Aldrich and N. Holonyak, Jr., *Proc. I.R.E.* **46,** 1236 (1958).
10. I. A. Lesk, *I.R.E. Transactions on Electron Devices,* ED-**6,** 28 (1959).
11. R. W. Aldrich and N. Holonyak, Jr., *J. Appl. Phys.* **30,** 1819 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 188–193. Printed in Great Britain

THE REGROWN-DIFFUSED TRANSISTOR

C. H. L. GOODMAN

Research Laboratories of the General Electric Company, Wembley, England

(*Received* 14 *March* 1960)

Abstract—A new process for producing high-frequency *n–p–n* silicon transistors is described. The novel feature is the formation of a doubly doped region in contact with singly doped material by the direct melting of appropriately doped fragments on to the latter, followed by subsequent solidification. A diffusion cycle then enables the required transistor structure to be developed.

The new technique is not limited to silicon and has a number of attractive features, among them: simplicity of apparatus, relatively high rate of production and efficient utilization of material.

Résumé—On donne un procédé nouveau de production de transistors au Silicium *n–p–n* à haute fréquence. La nouveauté est la formation d'une région doublement dopée avec un matérial simplement dopé par la fusion directe de fragments adéquatement dopés suivie d'une solidification. Un cycle de diffusion permet la réalisation de la structure recherchée du transistor.

Cette technique nouvelle n'est pas limitée au Silicium, et a un certain nombre d'avantages, parmi lesquels la simplicité des appareils, la vitesse relative de production et l'utilisation efficace du matériel.

Zusammenfassung—Es wird ein neuer Prozess zur Herstellung von Hochfrequenz-*n–p–n*-Transistoren aus Silizium beschrieben. Die Besonderheit besteht in der Bildung eines doppelt dotierten Bereiches auf einfach dotiertem Material durch direktes Aufschmelzen entsprechend dotierter Teile auf dieses Material und nachfolgendem Erstarren. Durch einen Diffusionszyklus erhält man dann die gewünschte Transistorstruktur.

Die neue Technik ist nicht auf Silizium beschränkt und hat ausserdem eine Anzahl attraktiver Merkmale, wie z.B. Einfachheit der Apparatur, relativ hohe Produktionsquote und gute Ausnutzung des Materials.

INTRODUCTION

A POSSIBLE method of making silicon transistors is by means of the melt-back diffused process.[1] In this process a material containing a high concentration of donor impurity and a somewhat lower concentration of acceptor impurity is partially melted and then allowed to recrystallize. The segregation of impurities that occurs on solidification leads to the formation of an *n*-type region with a relatively low impurity concentration [see Fig. 1(a)]. The acceptor impurity in the initial material is chosen to have a diffusion constant appreciably higher than that of the donor impurity, and with suitable heat treatment diffuses into the recrystallized region to produce a *p*-type base region [see Fig. 1(b)].

One of the major difficulties encountered with the above process is the preparation of single-crystal silicon, doubly doped with *p*- and *n*-type impurities to precise (and high) concentrations. However, it appeared that an inverted process using singly doped material, indicated in Fig. 2(a) and (b), had a number of interesting possibilities. In it, a heavily doped *n*-type region containing both donor and fast-diffusing acceptor impurities is produced in contact with lightly doped *n*-type silicon, by melting appropriately doped silicon fragments on to the purer material. As before, the transistor structure is developed by subsequent heat-treatment.

The new process[2], which will be described in greater detail below, has specific advantages over other methods for producing diffused-base silicon transistors, namely:

(1) Simple apparatus can be used.

(2) Production rate is relatively high.

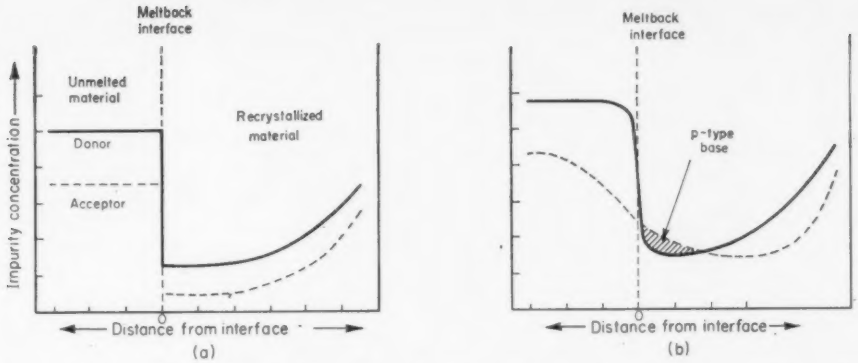


FIG. 1. (a) Impurity distribution after meltback. (b) Impurity distribution derived from that in (a) after diffusion cycle.

(3) A large area structure is produced which can be cut up to give bars of any required cross section.
(4) Base width can be accurately controlled down to 1 μ and possibly less.
(5) Base region is relatively flat.
(6) Only singly doped materials are required, and fairly wide resistivity ranges are acceptable.
(7) Efficient use is made of material (i.e. 500 transistors can be made from an initial 5 g of silicon).

THEORETICAL CONSIDERATIONS

The regrown-diffused process, like all techniques based on the simultaneous diffusion of acceptor and donor impurities, requires one impurity to diffuse appreciably faster than the other. (The known diffusion coefficients for impurities in silicon accordingly limit the process to the production of n–p–n structures.) In order that high emitter and base doping levels can be obtained, it is also desirable that the impurities segregate but little during regrowth. Table 1 shows some relevant data, obtained from the literature, on solid–liquid partition coefficient K and diffusion constant D in silicon for impurities of interest. The two requirements given above are best satisfied, it would appear, by boron as acceptor impurity and arsenic as donor. Aluminium and gallium could also be used as acceptor impurities, and antimony as donor, but the appreciable segregation of these elements occurring

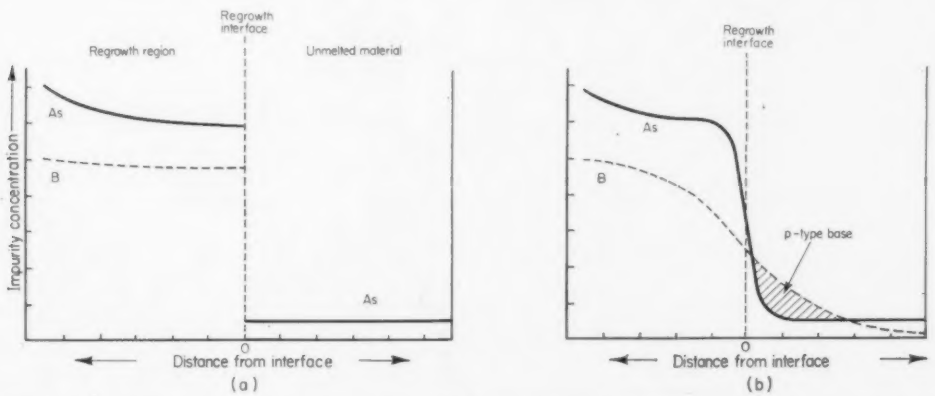


FIG. 2. (a) Impurity distribution after regrowth (b) Impurity distribution derived from that in (a) after diffusion cycle.

during regrowth could limit attainable doping levels.

The work reported here has been conducted almost entirely with boron and arsenic. The type of distribution to be expected with these impurities is indicated diagrammatically in Fig. 2(a) and (b). Segregation effects are small since K values are relatively close to unity. A detailed plot

Table 1

Element	Solid–liquid partition coefficient (K)	Diffusion constant (D) ($cm^2/sec \times 10^{11}$)	
		1100°	1200°
B	0·82	0·035	0·27
Al	0·004	0·16	1·2
Ga	0·01	0·053	0·39
In	0·0005	0·008	0·08
Tl	?	0·008	0·08
P	0·35	0·035	0·27
As	0·3	0·0033	0·025
Sb	0·04	0·0019	0·020
Bi	?	?	0·017

of net carrier concentration near the base region (assuming 100 per cent ionization), calculated for a diffusion time of 25·7 hr at a temperature of 1200°C,* is shown in Fig. 3. The different curves show the effect of varying the arsenic-to-boron concentration ratio in the doubly doped region prior to diffusion.

The following assumptions are implicit in this figure:

(i) The doubly and singly doped regions are of infinite extent and, before diffusion, have constant impurity concentration.

(ii) Diffusion constants and K values are those given in Table 1.

(iii) The lightly doped n-type region (collector) has an impurity concentration one hundredth that of the boron in the doubly doped region.

The first assumption appears valid in view of the small distances involved. Assumption (ii), however, is more open to doubt, but is necessary on account of the absence of more precise data.†

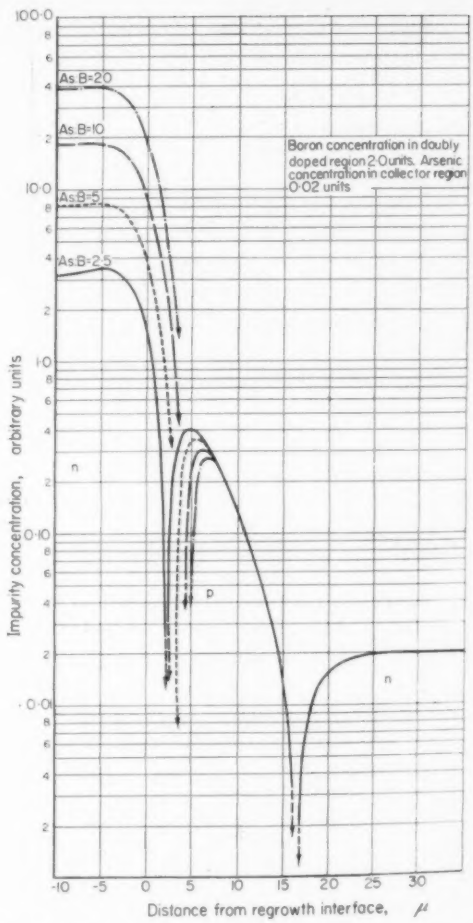


FIG. 3. Net carrier concentration in the regrown-diffused silicon transistor for various As–B ratios. Diffusion schedule 25·7 hr at 1200°C.

Fig. 4 shows the variation of base width and net maximum base-doping level with arsenic-to-

* This time and temperature simplify the diffusion calculations.

† It should be noted that base widths shown up by selective staining (in HF+1 per cent HNO_3, under strong illumination) are invariably appreciably less than those calculated theoretically. This could be explained by: (a) marked non-equilibrium segregation of As, which seems improbable, (b) a reduction in the assumed difference between the diffusion constants of As and B due to concentration or interaction effects and (c) some inherent selective action of the stain. A preliminary study suggests that both (b) and (c) are concerned.

boron concentration ratio. It can be seen that both parameters are relatively little affected by changes in the latter.

Smaller base widths than those indicated would be obtained for shorter times or lower temperatures of diffusion, base width being proportional to (time)$^{1/2}$ or (D_{boron})$^{1/2}$ (to a good approximation base width is determined by the diffusion of boron only).

The impurity concentration in the singly doped collector region also affects base width. In Fig. 3 this concentration was assumed to be 1 per cent of the boron concentration in the doubly doped region, which represents a convenient value for devices of practical interest. The base width increases as the collector impurity concentration is reduced, although it should be noted that under normal operating conditions (collector voltage applied) the effective base width would not increase to the same extent.

Somewhat different impurity distributions are obtained in the case of other pairs of impurities, although, as already mentioned, difficulties can arise on account of marked segregation. It is interesting to note that some adjustment of the net impurity distribution for the arsenic + boron combination is possible if a proportion of antimony or aluminium is added. The effect of antimony is to sharpen the emitter junction, while aluminium largely serves to increase the grading of the collector junction (see Fig. 5).

It should be noted that the impurity distribution obtained by the regrown-diffused process is identical to that given by the grown-diffused process.[3] (In the latter process a singly-doped–doubly-doped interface is produced by pulling a crystal from a melt, lightly doped with donor impurity, into which a double doping charge is introduced during growth. Diffusion across the interface occurs during continued growth, giving the required structure). In terms of simplicity of apparatus, utilization of material and speed of production the regrown-diffused process appears to offer considerable advantages.

Finally it is of course clear that the use of the regrown-diffused process is not limited to silicon. The production of *p–n–p* high-frequency germanium transistors is a possibility, but, more generally,

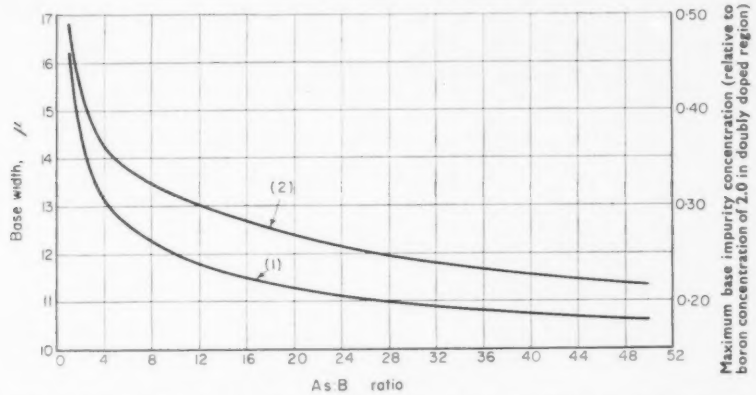


FIG. 4. (1) Variation of base width with As–B ratio. (2) Variation of net maximum base-impurity concentration with As–B ratio. (Diffusion schedule 25·7 hr at 1200°C.)

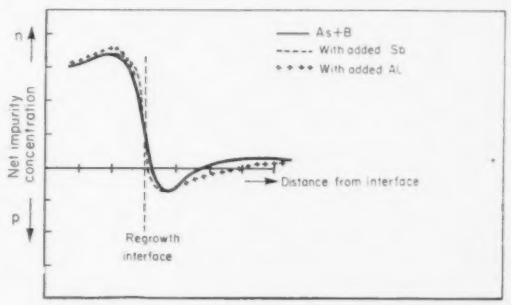


FIG. 5. Effect on impurity distribution of addition of Sb or Al (not to scale).

the process should be applicable to any semiconductor which can be melted without appreciable decomposition. It is also interesting to note that regrowth without diffusion enables alloy-like abrupt junctions to be prepared in silicon or germanium using impurities such as boron or phosphorus which cannot be alloyed directly.

EXPERIMENTAL PROCEDURES

Techniques and materials can be summarized under the following headings.

Starting materials

Boron- and arsenic-doped silicon of 25–60 and 1·0–7·0 mΩ-cm resistivity respectively were used in appropriate amounts to form the doubly doped region. The collector material was 1·5–2·0 Ω-cm *n*-type silicon. All materials were carefully etched, washed and dried before use.

Regrowth apparatus

A simple, but effective, apparatus is indicated diagrammatically in Fig. 6. A circulation of very pure dry hydrogen is maintained in the bell-jar. Melting is achieved by running the radiation heater at approximately 2700°C. Slight contamination of the melt with tungsten has been observed in some cases, though this had no apparent effect on electrical properties. (The tungsten appeared to segregate to the last part to freeze, forming tungsten silicide there.) The heater is fed from a variable voltage supply. The use of a specimen support of which the height can be varied during a run is helpful but not essential.

Regrowth technique

The aim of the technique adopted is to melt fragments of both types of doping material on to a block of collector material. The latter is conveniently 7–15 mm square.

The presence of any appreciable thickness of oxide on the collector or doping materials, or of traces of oxygen or water vapour in the gas stream, can lead to difficulties in getting the doped melt to wet the collector material, and also to "spitting" of the melt with consequent damage to the heater (which is in close proximity).

The depth of penetration of the melt into the collector material is controlled by observing the position of the liquid–solid interface, which is

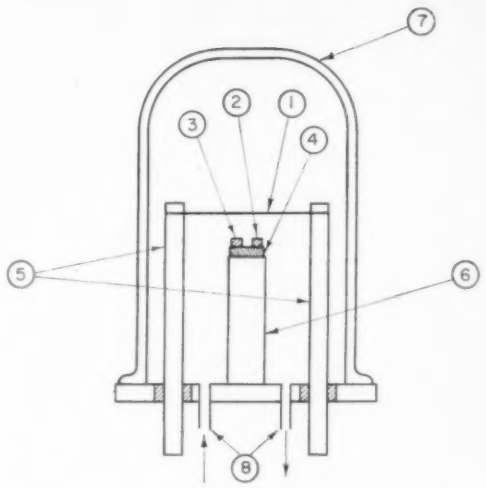


FIG. 6. Apparatus for producing regrown structures. (1) Tungsten heater strip. (2) Heavily doped *n*-type Si. (3) Heavily doped *p*-type Si. (4) Lightly doped *n*-type Si. (5) Water-cooled insulated pillars. (6) Water-cooled support. (7) Pyrex bell-jar. (8) Gas supply.

readily visible through a low-power stereo-microscope. When the required depth is reached, power is reduced and the silicon allowed to solidify slowly. The marked expansion of silicon on freezing and the contact angle between liquid and solid silicon result in the formation of a conical spike. A typical regrown specimen is shown in Fig. 7. Similar shapes have been reported by BILLIG and GASSON[4].

It is possible to obtain an extremely flat regrowth interface. Variations from flatness of less than $\pm 10\,\mu$ over distances of the order of 5 mm have been observed with 7 mm square collector blocks.

Diffusion

This has been carried out in sealed evacuated silica tubes for a time appropriate to develop the required structure, the specimens being withdrawn slowly from the furnace after treatment.

Transistor characteristics

The regrown-diffused specimens are cut up to give small bars. A thin aluminium wire is alloyed into the base region, and gold–antimony alloy contacts made to the emitter and collector regions.[3]

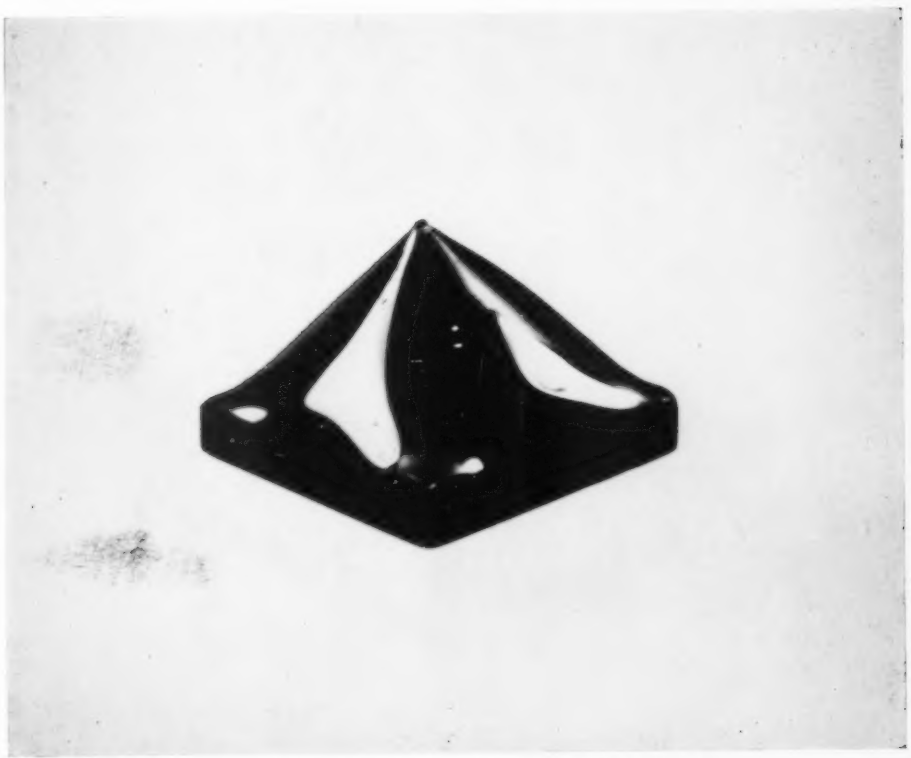


FIG. 7. A typical regrown specimen. (Base 8mm square.)

[*facing p.* 192

Current gains in the range 20–50 and cut-off frequencies (f_1) of 15–40 Mc/s are readily obtainable, while experimental devices with gains > 100 and f_1 up to 80 Mc/s have been produced. Unless the correct thermal conditions obtain during regrowth and diffusion cycles, minority-carrier lifetime is low and this limits the gain. However, lifetimes as good as those for the grown-diffused process can in fact be obtained.

Acknowledgements—It is a pleasure to record the help of Dr. T. R. NEILL, who has been entirely responsible for mounting and encapsulating the transistors and measuring their characteristics.

The author is indebted to the Admiralty for permission to publish this paper.

REFERENCES

1. A. B. PHILLIPS and A. M. INTRATOR, *I.R.E. National Convention Record* **5,** Part 3, 3–13, March, 1957. Institute of Radio Engineers, New York (1957).
2. *Brit. Pat. Appl.* No. 35903/58.
3. See for example: J. T. KENDALL, *Electron. Radio Engr.* **35,** 202 (1958).
4. E. BILLIG and D. GASSON, *J. Appl. Phys.* **28,** 1242 (1957).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 194–201. Printed in Great Britain

VARIATION WITH TEMPERATURE OF THE DISTRIBUTION COEFFICIENT OF INDIUM IN GERMANIUM

M. A. LEE

Research Laboratories of The General Electric Company, Wembley, England

(*Received* 28 *March* 1960)

Abstract—A relationship is derived for an alloyed-junction transistor giving the solid solubility of the emitter material (indium) in the base crystal (germanium) in terms of the slope of a plot of reciprocal current gain vs. emitter current. Measurements were carried out on batches of transistors alloyed at temperatures in the range 350–805°C. The form of the temperature variation of the distribution coefficient deduced from these measurements is in good agreement with that predicted by HALL down to about 550°C. Below this temperature the observed values of the distribution coefficient are higher than would be expected from HALL's relationship. A discussion of the possible causes of this divergence is presented.

Résumé—On déduit une équation, pour un transistor de jonction par alliage, qui donne la solubilité dans le solide du matériel émetteur (Indium) dans le cristal de base (Germanium) comme la pente de la courbe de l'inverse du gain de courant en fonction du courant émetteur. Des mesures ont été faites sur des transistors dont l'alliage est réalisé à des températures entre 350 et 805°C. La variation du coefficient de distribution déduite de ces mesures est en bon accord avec ceux donnés par HALL jusqu'à environ 550 °C. En dessous de cette température les valuers du coefficient de distribution sont supérieures à celles que l'on déduit de la relation de HALL. On discute les causes possible de ces divergences.

Zusammenfassung—Es wird eine Beziehung abgeleitet für einen Legierungstransistor, die die Löslichkeit des Emitter-Materials (Indium) im Basiskristall (Germanium) in Abhängigkeit von der Steigung der reziproken Stromverstärkung mit dem Emitterstrom ergibt. Es wurden Messungen an einer grossen Anzahl von Transistoren durchgeführt, die im Temperaturbereich von 350°C bis 805°C legiert worden waren. Die sich aus diesen Messungen ergebende Temperaturabhängigkeit des Verteilungskoeffizienten stimmt bis zu 550°C hinunter gut mit der überein, die HALL vorausgesagt hat. Für tiefere Temperaturen sind die beobachteten Werte des Verteilungskoeffizienten höher als nach HALL zu erwarten wäre. Die möglichen Ursachen dieser Abweichung werden diskutiert.

1. INTRODUCTION

THERE have been conflicting opinions regarding the behaviour, with temperature, of the distribution coefficient of indium in germanium. THURMOND *et al.*[1] found from measurements of the solute concentration in germanium crystals precipitated from slowly cooled melts that the distribution coefficient of indium in germanium varied little with temperature. By observing a systematic trend in the behaviour of different solutes in one solvent HALL[2] was able to predict that the distribution coefficient of indium in germanium should decrease rapidly with temperature, the variation being linear when plotted on a log (distribution coefficient) vs. reciprocal temperature basis. HALL also presented two experimental points obtained from measurements on germanium crystals grown from indium-rich solutions; these points indicated that the distribution coefficient did vary rapidly with temperature, but did not form sufficient evidence to confirm the prediction of a linear relation.

The experimental results of both THURMOND *et al.* and HALL are subject to the possibility of solute inclusions in the grown germanium crystals, whereas the results presented here are not. The

form of the temperature variation of the distribution coefficient is derived here from measurements of the electrical characteristics of transistors alloyed in the temperature range 805–350°C. These derived results are in good agreement with HALL's predicted linear variation down to about 550°C; below this the distribution coefficient appears to vary less with temperature.

2. THE DISTRIBUTION COEFFICIENT

In an alloyed-junction transistor the impurity concentration in the emitter recrystallized layer depends on the solid solubility of the emitter material in the base crystal at the alloying temperature. In particular, the case of a *p–n–p* transistor with an indium emitter and germanium base will be discussed. Relative values of impurity concentration can be calculated from the electrical characteristics of transistors alloyed at different temperatures, and from these values the temperature dependence of the solid solubility, and hence distribution coefficient, can be determined.

The distribution coefficient k can be related to transistor characteristics by considering the curve of reciprocal current gain α_{cb}^{-1} vs. emitter current I_e. The equation of this curve for a *p–n–p* transistor, due to WEBSTER[3] and modified by MISAWA[4] and MATZ[5], indicates a linear portion at high emitter currents having a slope*

$$\frac{\partial}{\partial I_e}\left(\frac{1}{\alpha_{cb}}\right) = \frac{\mu_{nb}W^2}{2D_{pb}A_e}\cdot\frac{\mu_{ne}\mu_{pe}}{\mu_{nb}\mu_{pb}}\cdot\frac{1}{\sigma_{pe}L_{ne}} \qquad (1)$$

where W is the base width and A_e is the effective area for current flow; the first subscript of the mobility μ, conductivity σ, diffusion constant D and diffusion length L refers to the carrier type; the second subscript refers to the region involved, emitter or base.

It is assumed in the derivation of this equation that the drift field introduced by the impurity-concentration gradient in the emitter recrystallized layer does not affect the diffusive nature of the minority-carrier transport in the emitter. This can be shown to be reasonable by a calculation of the relative velocities for drift and diffusion. A further assumption implicit in the equation is that the diffusion length of electrons in the emitter (L_{ne}) is less than the thickness of the emitter recrystallized layer, W_e. At higher alloying temperatures this assumption is valid, but since W_e decreases rapidly with decreasing alloying temperature, and L_{ne} changes very slowly with alloying temperature, L_{ne} becomes equal to W_e at about 400°C. Below this temperature the assumption is doubtful, and this point is discussed further in Section 5. It is difficult to predict the exact dependence of L_{ne} on alloying temperature as it involves mobility, lifetime and geometrical factors, but it is reasonable to assume that it is considerably less than W_e and in the opposite direction.

The last two factors in equation (1) contain all the terms dependent on acceptor concentration (N_A) and may conveniently be set equal to the reciprocal of a function $\Lambda(N_A)$. This function Λ then becomes identical with the factor regarded as $\sigma_{pe}L_{ne}$ in previous papers[3,7] ($\sigma_e L_e$ in their terminology) and is

$$\Lambda = \sigma_{pe}L_{ne}\frac{\mu_{nb}\mu_{pb}}{\mu_{ne}\mu_{pe}} \qquad (2)$$

This may be expressed in more fundamental terms by the use of $L_{ne}^2 = D_{ne}\tau_{ne}$ and the conductivity and Einstein relations giving

$$\Lambda = e\mu_{nb}\mu_{pb}\left(\frac{kT}{e}\right)^{1/2}\tau_{ne}^{1/2}\mu_{ne}^{-1/2}p_e \qquad (3)$$

where D_{ne} and τ_{ne} are the minority-carrier diffusion constant and lifetime in the emitter respectively, and p_e is the hole density in the emitter. Only the last three terms in this equation depend on N_A; the nature of their dependence is now examined in turn.

According to the HALL–SHOCKLEY–READ recombination theory[6,9] the minority-carrier lifetime is constant in the range of impurity concentration under consideration. The assumption is made here that alloying at different temperatures does not alter the concentration of recombination centres in the recrystallized layer.

The variation of electron mobility with acceptor concentration in *p*-type germanium (μ_{ne}) has been measured by PRINCE[10] for concentrations up to 2×10^{17} atoms/cm^3; his results are shown by the full

* In WEBSTER's original equation the base recombination term was considered to be dependent on emitter current; however, since the recombination is essentially a monomolecular process,[6] rather than a bimolecular process, this would appear to be incorrect.[7,8]

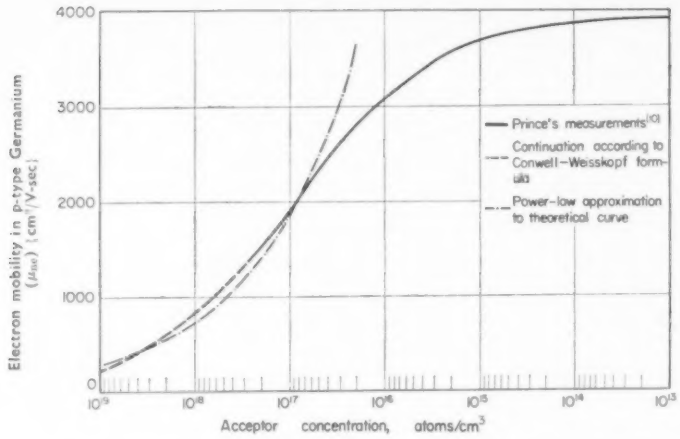


FIG. 1. The variation with acceptor concentration of electron mobility in p-type germanium.

line on Fig. 1. At higher acceptor concentrations PRINCE's results were found to be in accord with the CONWELL–WEISSKOPF[11] formula for impurity scattering. Theoretical values of μ_{ne} corresponding to the CONWELL–WEISSKOPF formula are shown as a broken-line curve continuing from PRINCE's results. In the acceptor concentration range 2×10^{16}–10^{19} atoms/cm³ this curve can be approximated by a power law (shown as the chain-line curve)

$$\mu_{ne} = aN_A^{-b} \tag{4}$$

where $a = 2{\cdot}63\times10^{10}$; $b = 0{\cdot}42$. The agreement between the power law and the theoretical curve is within 30 per cent in the acceptor concentration range 2×10^{16}–10^{19} atoms/cm³ which is adequate since μ_{ne} occurs as a fractional power in the equation.

In the emitter, the hole density is equal to the acceptor concentration if the acceptors are completely ionized. For the case of indium in germanium the acceptors are approximately 90 per cent ionized for $N_A = 10^{19}$ atoms/cm³ and rapidly become fully ionized as N_A is decreased. It is reasonable to assume that $N_A < 10^{19}$ atoms/cm³ and so the hole density may be set equal to the acceptor concentration.

Incorporating these considerations into the expression for Λ gives

$$\Lambda = e\mu_{nb}\mu_{pb}\left(\frac{kT}{e}\right)^{1/2}\tau_{ne}{}^{1/2}(aN_A{}^{-b})^{-1/2}N_A \tag{5}$$

The terms in this expression for Λ which are independent of N_A can be grouped together as c giving

$$\Lambda = cN_A{}^{1+b/2} \tag{6}$$

N_A in this equation is the acceptor concentration in the part of the recrystallized layer which corresponds to the edge of the emitter-junction space-charge region at any given bias, since it is in this part of the layer that holes are in equilibrium with holes in the base. Any change from N_A of the acceptor concentration deeper in the emitter recrystallized layer is unlikely to affect this equilibrium.

The quantity relevant to the determination of the distribution coefficient, however, is the concentration of acceptors at the edge of the recrystallized layer. This may differ slightly from N_A as defined above, but the discrepancy is likely to be negligible in view of the small thickness of the space-charge region on the emitter side. This space-charge region, which increases with decreasing acceptor concentration, is $0{\cdot}004\mu$ thick for the lowest concentration in the present case which is almost three orders of magnitude smaller than the thinnest recrystallized layer involved in the experiments ($2{\cdot}5\mu$). In this calculation $N_A = 2\times10^{17}$ atoms/cm³, the donor concentration in the base is 2×10^{15} atoms/cm³ and it is assumed that the residual potential across the junction is 0·2 V.

The part of the layer under consideration recrystallizes at, or very slightly below, the alloying

temperature and will have dissolved in it an amount of indium determined by the solid solubility (x_s) at the alloying temperature given by

$$x_s = \frac{N_A}{N_c} = \left(\frac{\Lambda}{c}\right)^{1/(1+b/2)} \frac{1}{N_c} \qquad (7)$$

N_c being the total density of germanium atoms in the base crystal. The distribution coefficient is $k = x_s/x_L$, where x_L is the atomic fraction of indium in the liquid phase. Converting to logarithmic form and using the value of x_s given by equation (7) one obtains:

$$\log k = \left(\frac{1}{1+b/2}\right) \log \Lambda - \log x_L + \text{constant} \qquad (8)$$

Λ multiplied by a geometrical factor, which is at this stage unevaluated, is calculated from equation (1); this geometrical factor is absorbed in the constant of equation (8) which is discussed subsequently in Section 4. x_L is known from the liquidus of KECK and BRODER[12] and KLEMM *et al.*[13]

3. EXPERIMENTAL PROCEDURE

Batches of transistors were alloyed at temperatures in the range 350–805°C. Each batch was made from germanium base wafers of the same thickness on to which indium emitters and collectors were simultaneously alloyed at a temperature T°C. The indium was used in the form of spheres and stated by the manufacturers to be of 99·97 per cent purity; no analysis of these spheres was made here. The batch wafer thicknesses were chosen so that the mean values of base width were approximately the same for all batches, in order to maintain a similar geometrical structure. From each batch a group of four to ten transistors, having base widths lying in the range ± 25 per cent of the mean value, were selected for measurement. The base widths were calculated from the common-base cut-off frequency $f_{\alpha}{}^{ce}$ using the relation

$$f_{\alpha}{}^{ce} = \frac{\pi D_{pb}}{W^2} \qquad (9)$$

The current gain to be measured is defined as the small-signal common-emitter short-circuit current gain (α_{cb}). This was measured using a bridge circuit arranged so that the emitter bias current was supplied in the form of pulses in order to avoid thermal effects.

A_e, defined as the effective area for current flow, was taken as the area of the emitter bead, after WEBSTER[3].

4. RESULTS

Mean Λ values for each group of transistors, calculated from the slope of reciprocal current gain vs. emitter current curves using equation (1), are plotted logarithmically against $10^3/T$ in Fig. 2. The

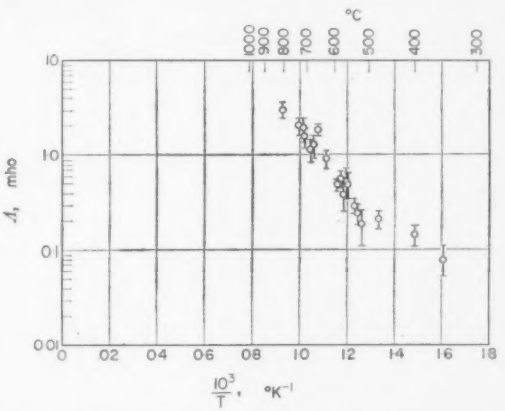


FIG. 2. The temperature variation of Λ. [Λ is defined by equation (2).]

standard deviation of the Λ values for each group is represented by the extension of the points along the Λ axis. Equation (8) is used to translate these Λ values into relative values of the distribution coefficient of indium in germanium which, when plotted as log k vs. $10^3/T$ (Fig. 3), give a straight line at higher temperatures and diverge below about 550°C. By assuming that this straight line passes through the well-established value of the distribution coefficient at the melting point of germanium (T_m, k_m) the relative k values can be converted into absolute values. In this way the constant in equation (8) is adjusted through a knowledge of the point $M(T_m, k_m)$.

The variation with temperature of the solid solubility of indium in germanium is shown in Fig. 4. This is calculated from the k values of Fig. 3 using $x_s = kx_L$, x_L being given by the liquidus of KECK and BRODER. x_s is calculated in this way, rather than directly from equation (7), since the

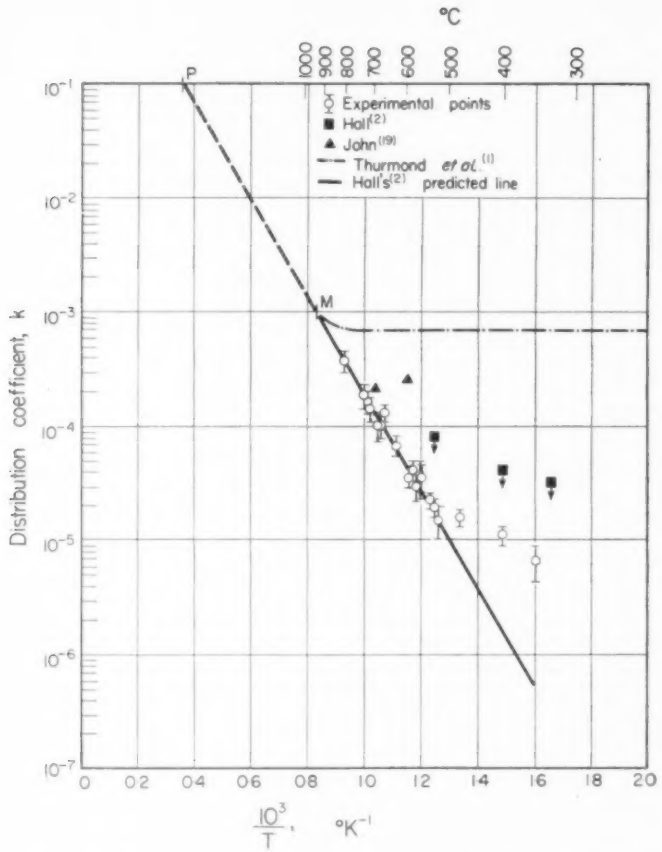


FIG. 3. Variation with temperature of the distribution coefficient of indium in germanium.

constant in equation (8) is more easily determined than that in equation (7).

KECK and BRODER's liquidus is shown in Fig. 4. The ideal liquidus also shown is calculated from the equation

$$\ln(1-x_L) = -\frac{\Delta H^f}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right) \qquad (10)$$

where ΔH^f is the latent heat of fusion of the pure solvent taken as 7700 cal/mole[14], T_m is 938°C[15,16] and R is the gas constant 1·986 cal/mole per deg. The experimental and ideal liquidus curves are seen to be very similar (Fig. 4).

5. DISCUSSION

The form of the temperature variation of the distribution coefficient of indium in germanium deduced here from transistor characteristics is in good agreement at higher temperatures with HALL's predicted line[2] (Fig. 3). This predicted line passes through $P(T = 2{\cdot}3T_m,\ k = 0{\cdot}1)$, a point common to all slightly soluble solutes in germanium, and $M(T_m, k_m)$ the distribution coefficient at the melting point: the latter point M is known from radioactive-tracer and conductivity measurements.[17] It is apparent that the form of the variation presented here differs substantially from that reported by THURMOND *et al.*[1] and does

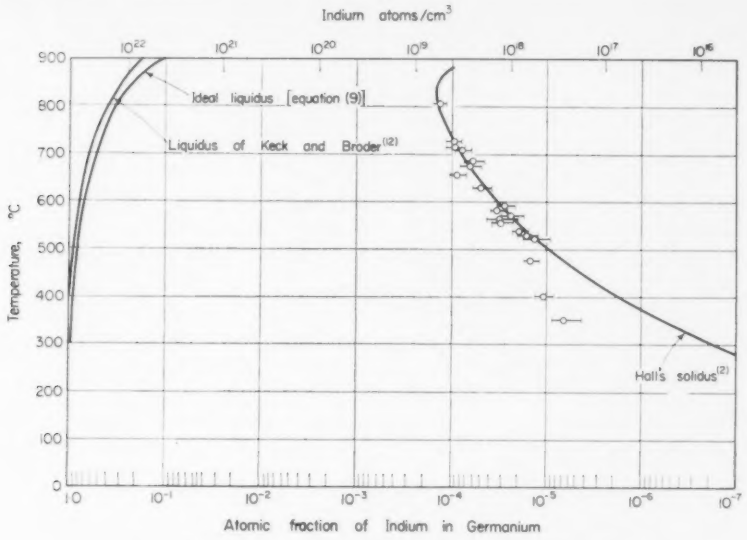


FIG. 4. The solid solubility of indium in germanium.

not corroborate the suggestion recently put forward by TRUMBORE[18]. This suggestion is that the experimental results presented by HALL[2], JOHN[19] and TRUMBORE[18] are not upper limits of the solid solubility as proposed by HALL, but represent reasonably good estimates of the true solid solubility.

At temperatures below about 550°C the results obtained here diverge from HALL's predicted linear relationship. This divergence could be attributed to:

(1) An assumption, used in deducing distribution coefficient values from transistor characteristics, becoming invalid.
(2) The presence of p-type impurities other than indium in the emitter recrystallized layer.
(3) Non-equilibrium conditions in the liquid phase at the onset of recrystallization.
(4) The non-ideality in either the liquid or solid phase.

Discussing these interpretations in turn: two assumptions were made in deriving a relation between Λ and the distribution coefficient, firstly that the diffusion length of electrons in the emitter (L_{ne}) was less than the depth of the recrystallized layer (W_e), and secondly that the lifetime of electrons in the emitter did not depend on the alloying temperature. Regarding the first assumption that $L_{ne} < W_e$, it was mentioned earlier (Section 2) that this appears to be questionable for transistors alloyed at temperatures below 400°C. However, the violation of this assumption could only account for smaller values of the distribution coefficient at lower temperatures rather than the observed larger values. This follows, since L_{ne} in the expression for Λ [equation (2)] would have to be replaced by the smaller quantity W_e, giving a lower Λ value, and hence reducing the distribution coefficient. The second assumption concerns the effective lifetime of electrons in the emitter (τ_{ne}); this is affected by the presence in the emitter recrystallized layer of crystal imperfections and by the presence of impurities having energy levels in the centre of the forbidden gap. Copper and nickel are typical of such impurities, but as they have very low distribution coefficients in germanium ($\sim 10^{-10}$ at 550°C) they will have negligible solid solubility in the recrystallized layer. There would not appear to be any reason for the concentration in the recrystallized layer of either impurities or crystal imperfections to change abruptly at 550°C, consequently τ_{ne} would not be expected to change in the region of 550°C.

The observed divergence of the distribution coefficient from linearity in the temperature range 550–350°C would be accounted for by the presence

of an additional impurity in the recrystallized germanium in a concentration of 7 p.p.m. *p*-Type impurities such as gallium, aluminium or zinc having large binary distribution coefficients in germanium (10^{-1}–10^{-2} varying little with temperature)[18,20] would have high solid solubilities in the recrystallized germanium. These particular additional impurities are quoted by the manufacturers as being present in their indium spheres in concentrations less than 10 p.p.m.; this would introduce an impurity concentration of $\sim 0{\cdot}1$ p.p.m. into the recrystallized germanium for each additional impurity (this impurity concentration is calculated using $k = 10^{-2}$[20] for the distribution coefficients of additional impurities in the temperature range of interest). It is incorrect, however, to use values of distribution coefficients which are relevant to binary systems for the multiple component system under consideration. It is likely that the distribution coefficients of additional impurities in a multiple-component system may differ appreciably from their binary values, and consequently the impurity concentration in the recrystallized germanium due to additional impurities may differ considerably from the value of $\sim 0{\cdot}1$ p.p.m. stated above. The element boron has not been included in this discussion because although it has a very high binary distribution coefficient in germanium ($k = 17$)[21] and would have to be present in a concentration of only $\sim 0{\cdot}4$ p.p.m. in the liquid alloy to account for the observed non-linearity, it appears to be unlikely that boron is present as it has not been found possible even to introduce it deliberately due to very effective removal by oxidation.[22] This discussion shows that the possibility of additional impurities causing the observed departure from linearity cannot be dismissed.

Equilibrium conditions may be considered to exist during recrystallization if the rate at which the alloy cools is sufficiently slow. It is only the acceptor concentration in the initial part of the recrystallized layer which is of interest in this context, as was discussed in Section 2. The initial part of the layer recrystallized at a rate which did not exceed 6°C/min, and was probably considerably less; this represents a very low growth rate.

The temperature variation of the distribution coefficient suggested by HALL was $\log k = B - A/T$, where A and B are independent of temperature. This followed as a restricted case from a more general equation due to THURMOND and STRUTHERS[23] who included a term $\log \gamma^l_{\mathrm{In}}$ in the expression for $\log k$, γ^l_{In} being the activity coefficient of indium in the liquid phase, which could produce curvature on a plot of $\log k$ vs. $10^3/T$ at lower temperatures. Values of γ^l_{In} were found here for the In–Ge system by first deriving the activity coefficient of germanium in the liquid phase from KECK and BRODER's liquidus, and then calculating γ^l_{In} using the DUHEM–MARGULES equation[24]: these calculated values were very close to unity in the temperature range of interest (350–550°C) and would consequently only very slightly affect the linearity of the $\log k$ vs. $1/T$ plot. ORIANI and HALL[25] point out that the expression for $\log k$ should also contain a further term involving heat-capacity differences. These differences appear to be rather small[26] in the temperature range of interest and likely only to introduce a very small term in the expression for $\log k$.

6. CONCLUDING REMARKS

The form of the temperature variation of the distribution coefficient of indium in germanium which has been deduced here from electrical measurements on transistors tends to confirm HALL's prediction. From the foregoing discussion it appears that the observed divergence from linearity in the $\log k$ vs. $10^3/T$ plot at low temperatures is probably due to the presence of additional impurities.

REFERENCES

1. C. D. THURMOND, F. A. TRUMBORE and M. KOWALCHIK, *J. Chem. Phys.* **25,** 799 (1956).
2. R. N. HALL, *J. Phys. Chem. Solids* **3,** 63 (1957).
3. W. M. WEBSTER, *Proc. I.R.E.* **42,** 914 (1954).
4. T. MISAWA, *J. Phys. Soc. Japan* **10,** 362 (1955).
5. A. W. MATZ, *Proc. I.R.E.* **46,** 616 (1958).
6. W. SHOCKLEY and W. T. READ, *Phys. Rev.* **87,** 835 (1952).
7. L. D. ARMSTRONG, C. L. CARLSON and M. BENTIVEGNA, *R.C.A. Rev.* **17,** 37 (1956).
8. N. I. MEYER, *J. Electronics and Control* **4,** 305 (1958).
9. R. N. HALL, *Phys. Rev.* **87,** 387 (1952).
10. M. B. PRINCE, *Phys. Rev.* **92,** 681 (1953).
11. E. CONWELL and V. F. WEISSKOPF, *Phys. Rev.* **77,** 388 (1950).
12. P. H. KECK and J. BRODER, *Phys. Rev.* **90,** 521 (1953).

13. W. KLEMM, L. KLEMM, E. HOFFMANN, H. VOLK, E. ORLAMUNDER and H. KLEIN, *Z. anorg. Chem.* **256,** 239 (1948).
14. O. KUBASCHEWSKI and E. EVANS, *Metallurgical Thermochemistry* (2nd Ed.). Wiley, New York (1956).
15. H. STÖHR and W. KLEMM, *Z. anorg. u. allgem. Chem.* **241,** 305 (1939).
16. F. X. HASSION, C. D. THURMOND and F. A. TRUMBORE, *J. Phys. Chem.* **59,** 1076 (1955).
17. J. A. BURTON, *Physica, 's Grav.* **20,** 845 (1954).
18. F. A. TRUMBORE, *Bell Syst. Tech. J.* **39,** 205 (1960).
19. H. F. JOHN, *J. Electrochem. Soc.* **105,** 741 (1958).
20. F. A. TRUMBORE, E. M. PORBANSKY and A. A. TARTAGLIA, *J. Phys. Chem. Solids* **11,** 239 (1959).
21. H. E. BRIDGERS and E. D. KOLB, *J. Chem. Phys.* **25,** 648 (1956).
22. The author is indebted to the referee for this information.
23. C. D. THURMOND and J. D. STRUTHERS, *J. Phys. Chem.* **57,** 831 (1953).
24. P. DUHEM and M. MARGULES, (p. 63 of Ref. 14).
25. R. A. ORIANI and R. N. HALL, *J. Phys. Chem. Solids* **6,** 97 (1958).
26. K. K. KELLEY, *Bulletin of the U.S. Bureau of Mines* No. 476 (1949).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 202–210. Printed in Great Britain

CROSS-SECTIONAL RESISTIVITY VARIATIONS IN GERMANIUM SINGLE CRYSTALS

J. A. M. DIKHOFF

Philips Research Laboratories, N. V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands

(*Received* 22 *March* 1960; *in revised form* 20 *May* 1960)

Abstract—Germanium single crystals frequently show considerable cross-sectional resistivity variations. The dependence of these variations upon crystal orientation as well as upon the shape of the solid–liquid interface during crystal growth has been investigated. It appears that a direct relationship exists between the crystal orientation, the shape of the growth interface and the occurrence of persistent cross-sectional resistivity variations. The origin of these variations is closely related to the presence of a flat facet on the otherwise curved growth interface, which facet coincides with a {111} plane. The segregation constants of several impurities appear to be different for the flat part and for the curved part of the growth interface, thus causing cross-sectional variations of the impurity concentration. Methods for preventing these variations are discussed.

Résumé—Les monocristaux de Germanium montrent des variations considérables de la résistivité. On étudie les relations de ces variations avec l'orientation du cristal aussi bien qu'avec la forme de l'interface solide–liquide pendant la croissance du cristal. On montre qu'il existe une relation directe entre l'orientation cristalline, la forme de l'interface de croissance et l'apparition des variations persistantes de la résistivité des profils. L'origine de ces variations est étroitement liée à la présence d'une face plane sur l'interface qui ordinairement devrait être courbée, cette face coincide avec le plan {111}. Le coefficient de ségrégation de plusieurs impuretés est différent pour la partie plane et pour la partie courbée de l'interface de croissance. Ceci donne des variations de profils dans la concentration d'impuretés. On donne des méthodes pour empêcher ces variations.

Zusammenfassung—Germanium-Einkristalle zeigen häufig über den Querschnitt beträchtliche Schwankungen des Widerstandes. Es wurde die Abhängigkeit dieser Widerstandsschwankungen sowohl von der Kristallorientierung als auch von der Form der Grenzfläche zwischen Schmelze und Kristall während des Kristallwachstums untersucht. Es zeigt sich, dass eine direkte Beziehung zwischen der Kristallorientierung, der Form der Erstarrungsfront und dem Auftreten von Widerstandsschwankungen über den Kristallquerschnitt besteht. Das Entstehen dieser Widerstandsschwankungen ist eng mit dem Vorhandensein einer flachen Facette auf der im übrigen gekrümmten Erstarrungsfläche verbunden. Diese Facette fällt mit einer {111}-Ebene zusammen. Die Verteilungskoeffizienten einiger Störstellen sind anscheinend für den flachen und den gekrümmten Teil der Erstarrungsfläche verschieden, wodurch sie Schwankungen der Störstellenkonzentration über den Querschnitt des Kristalles verursachen. Es werden Methoden zur Vermeidung dieser Schwankungen diskutiert.

I. INTRODUCTION

GERMANIUM single crystals for semiconductor use are required to have a controlled and constant electrical resistivity. In recent years, several methods of growing crystals which meet this requirement have been described, e.g. the horizontal zone-levelling technique (PFANN and OLSEN[1]), the programmed growth-rate technique (BURTON *et al.*[2]) and the floating-crucible technique (LEVERTON[3], GOORISSEN and KARSTENSEN[4]). A recent survey of these and other techniques has been given by OKKERSE[5].

Most attention has been paid to the distribution of impurities along the length of the crystal. However, little has been published on the cross-sectional distribution of impurities. CRESSEL and POWELL[6] have reported rather large resistivity variations across crystals grown with the horizontal zone-levelling technique. These variations were attributed to the inadequacies of thermal stirring in the melt. On the other hand BENNETT and SAWYER[7] did not find any serious variation in crystals grown with the same technique. GOSS and ADLINGTON[8] found transverse resistivity

variations in gallium-doped silicon crystals which had been pulled vertically with a small rotation speed. Such variations did not appear in crystals pulled with larger rotation speeds. The authors thought that this effect might be explained by the formation of a vortex in the melt when the growing crystal is rapidly rotated. CAMP[9] thought on the contrary, that a high rotation speed may produce stable flow patterns which might introduce transverse resistivity variations. Many other investigators have found that germanium crystals sometimes show considerable cross-sectional resistivity variations whose origin is not well understood.

Considerable cross-sectional resistivity variations are often observed in germanuim single crystals grown by the vertical-pulling technique. A typical plot of the resistivity as measured at room temperature with the four-point method along the diameter of a crystal pulled in the [111] direction and doped with antimony is given in Fig. 1. A

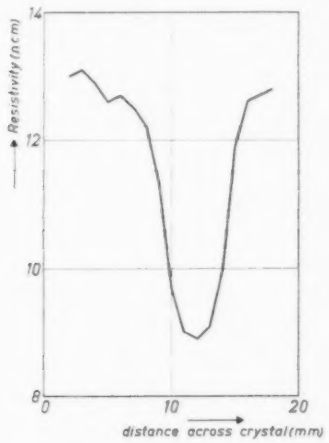


FIG. 1. Cross-sectional resistivity of an antimony-doped crystal, as measured with a four-point probe.

rather sharply defined region* of lower resistivity is found at or near the centre of the major part of the crystal. OKKERSE[10] found that the ratio of the highest resistivity and the lowest resistivity as measured with a four-point probe is more or less constant at 1·5 in pulled crystals doped with antimony.

* In this paper, this region of the crystal will be referred to as the "core" of the crystal.

Sometimes the centre of the core does not coincide with the centre of the crystal, the position relative to the crystal axis being the same over the whole length of the crystal. This fact already gives some information concerning the origin of cores. It does not seem very likely, for example, that cores originate from a stirring action in the melt. For, if the formation of an asymmetric core were caused by a stable flow pattern in the melt as suggested by CAMP[9], then one would have to assume that an eccentric flow pattern can be generated by, and at the same time rotate with, the symmetrically rotating crystal, and that the flow pattern should remain constant during the growth of the whole crystal.

Approaching the problem from the solid phase one might find other indications as to the origin of cores. After rapid withdrawal of a crystal from the melt, a flat region is sometimes found on the otherwise curved growth interface (GOORISSEN[11], WILKES[12]). This flat region is situated at that part of the interface where a ⟨111⟩ direction is approximately perpendicular to the interface, and is therefore thought to be a more or less perfect {111} plane. This observation prompted us to investigate the effect of the orientation of the seed crystal and the shape of the growth interface on the resistivity cores.

It will be shown in this paper that the occurrence of persistent resistivity variations across the crystal is indeed closely related to the shape of the solid–liquid interface during growth of the crystal. It also appears that the observed phenomena depend critically on the orientation of the crystal. It will be shown that even small changes in the angle between the [111] direction and the growth direction may cause considerable changes in the cross-sectional resistivity pattern. The effect of the nature of the added donor or acceptor impurities will also be discussed. Finally, several methods for preventing cross-sectional resistivity variations will be given.

II. EXPERIMENTAL

Germanium single crystals were grown from the melt using the vertical crystal-pulling technique. The growth apparatus is similar to that described by TEAL and LITTLE[13]. If not otherwise mentioned, crystals were pulled from a normal graphite crucible as shown schematically in Fig.

2(a); the crystals usually weighed about 125 g and had a diameter of about 20 mm. In one case a crucible as shown in Fig. 2(b) was used in order to obtain a less curved growth interface as was

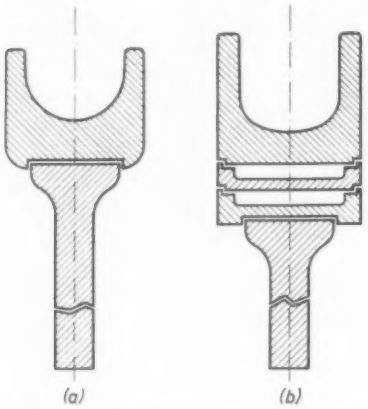


FIG. 2. (a) Normal crucible. (b) Crucible with underheater.

suggested by OKKERSE[10]. A few crystals were grown from a floating crucible as described by GOORISSEN and KARSTENSEN[4]. The thermal conditions in this crucible are such that the growth interface is very flat. All crystals were grown in an inert gas atmosphere consisting of 76 per cent N_2 and 24 per cent H_2. In most cases the pull rate was 1 mm/min and the rotation rate 45 rev/min.

A flat face parallel to the crystal axis was ground on the periphery of the crystals. The crystallographic orientation was determined by standard X-ray techniques. The [111] direction closest to the crystal axis is defined [see Fig. 3(a)] by the angle θ between the [111] direction, and the growth direction and the angle ϕ between the perpendicular to the ground face and the projection of the [111] direction on a plane perpendicular to the crystal axis.

The crystals were cut in slices perpendicular to the crystal axis each slice being about 5 mm thick. The resistivity* of each slice was measured as a function of distance along the diameter with a four-point probe (point spacing 0·9 mm). This was done in four directions which are related to the projection of the [111] direction as is shown in Fig. 3(b). In some cases the resistivity was also measured with the one-point method on thin bars because of the better resolution obtained by this method.

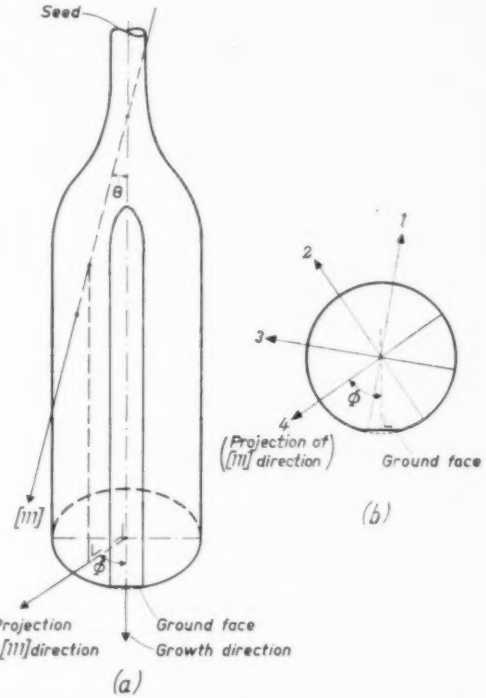


FIG. 3.(a) Diagram showing the relation between the external shape of a crystal and the [111] direction closest to its axis. (b) Directions in which the resistivity was measured on the cross-section.

One method of studying the solid–liquid interface during growth is to withdraw the growing crystal rapidly from the melt. However, part of the interface is always covered by a frozen drop which is retained on the interface by surface tension. Some of the liquid may also be retained by surface tension on other parts of the interface, thus masking a possible fine structure which might exist on the actual growth interface. Moreover, quite a number of crystals have to be grown in order to study the shape of the interface at different stages of growth. All this was felt to be a disadvantage of this method.

The shape of the growth interface can also be revealed by copper-plating of a grown p–n junction.

* The resistivity was always measured at room temperature.

This method was used for one crystal that had been grown with a number of p–n junctions by alternately adding adequate amounts of antimony or indium.

A third way to obtain valuable information on the shape of the interface during all stages of growth is to study the pattern of striations as revealed by a pulsed copper-plating technique described by CAMP[9] and modified for n–type germanium by BLOEM[14]. In the present investigation only n-type crystals were plated in an alkaline copper solution (consisting of 16 g NaOH, 1 g $CuSO_4 \cdot 5\,H_2O$, 19 g tartaric acid made up with deionized water to 100 ml), the germanium sample being the anode. The applied voltage was 1000 V and the pulse frequency was about 10 pulses/sec. The plating pattern is believed to be formed by preferentially etching and plating of regions with different resistivity. These regions result from variations in growth rate during the pulling of the crystal, which variations are probably due to an asymmetric temperature distribution in the melt as there is a one-to-one relation between the rotation speed and the number of the striations. Hence the crystal contains a helical surface of finite thickness in which the resistivity differs from that in the rest of the crystal. One "step" of the helical surface is formed during one revolution of the growing crystal. For a pull rate of 1 mm/min and a rotation rate of 45 rev/min the distance between two "steps" is about 25μ. Each "step" of the helical surface conforms to the shape of the growth interface, and the plating pattern on a longitudinal section thus indicates successive positions of the growth interface.

III. RESULTS

a. *Crystal orientation*

In order to investigate the effect of the crystal orientation upon the formation of cores, the following experiments were carried out. An antimony-doped crystal (crystal no. 1) was grown on a seed crystal in which the [111] direction made a small angle with the crystal axis. From X-ray measurements it appeared that $\theta = 7°$ and $\phi = 45°$ [see Fig. 3(a)]. The cross-sectional resistivity was measured in four directions as indicated in Fig. 3(b) on each of the ten slices cut from the crystal. In direction 4 of all slices, a large eccentric dip in the resistivity was found, the value of ρ_{max}/ρ_{min} varying from 1·39 to 1·53. In all slices this dip was situated at the same side of the crystal centre with an average distance from the centre of the crystal to the centre of the dip of 4 mm. For one typical slice, plots of resistivity versus distance are given in Fig. 4 for the four directions indicated. These data can also be presented in a resistivity contour map like the one shown in Fig. 5. The relation between the [111] direction in the crystal and the direction in which the eccentric resistivity dip is found [(see Fig. 3(a) and (b)] suggests that the origin of cores is correlated with the crystal orientation. Additional evidence was obtained from an experiment with an antimony-doped crystal in which θ was 13°. The core was again eccentric and related to the [111] direction in the same way. The

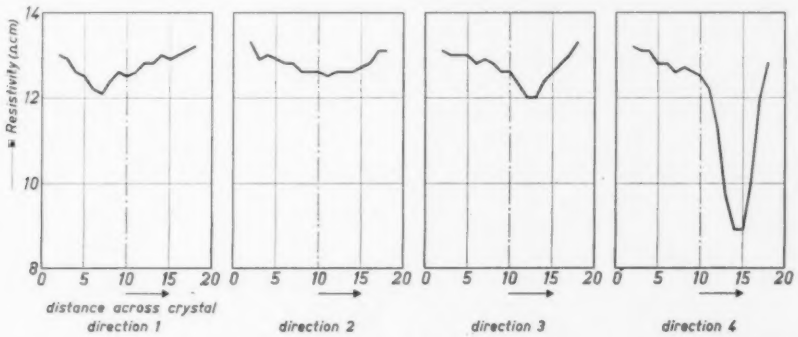


FIG. 4. Resistivity versus distance in four directions across a typical slice of crystal No. 1 (four-point measurement).

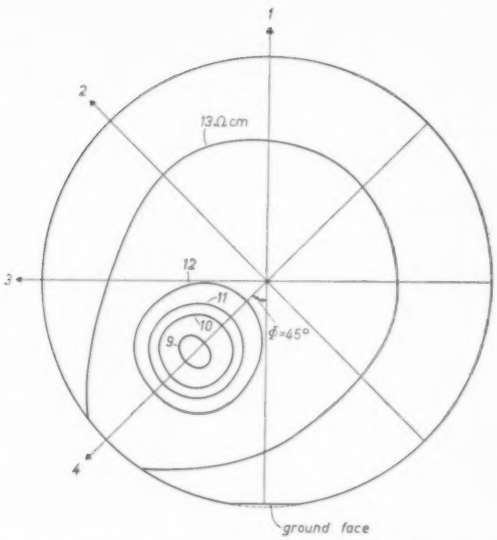


FIG. 5. Resistivity contour map of a typical slice from crystal no. 1 (four-point measurement).

average value of ρ_{max}/ρ_{min} was 1·47. However, in this case the distance between the centre of the core and the crystal axis was about 7 mm. Thus it appears that increasing the angle θ between the [111] direction and the growth direction moves the core to the outside of the crystal. In crystals grown in the [110] direction ($\theta = 35°$) and in the [100] direction ($\theta = 55°$) no resistivity dips have been found in any direction across the crystal. In these crystals only small cross-sectional resistivity variations were observed; the value of ρ_{max}/ρ_{min} was always smaller than 1·05. These experiments thus prove that a direct relationship exists between the occurrence of resistivity cores and the orientation of the crystal.

b. *Shape of the growth interface*

A peculiar feature of the crystals described in the preceding paragraph may be mentioned. It appeared that the distance from the centre of the core to the crystal axis is not constant over the whole crystal but becomes larger at the lower end of the crystal. It was thought that this effect might be connected with a change of the curvature of the solid–liquid interface when this approaches the bottom of the crucible. The shape of the growth interface was therefore studied.

The growth interface of a crystal with $\theta = 7°$ as exposed by rapid withdrawal of the growing crystal from the melt is shown in Fig. 6. The interface was found to be convex to the melt, with a flat facet (partly covered by a frozen drop) situated out of centre. This indicates that the freezing mechanism is not the same for every point of the growth interface.

For a closer examination of the shape of the growth interface a crystal with $\theta = 7°$ was grown with a number of p–n junctions. A slice was cut from an n-type region of the crystal and an eccentric core was found on the cross-section of this slice, the position of the core being correlated with the crystal orientation as described before. A part of the crystal containing a p–n junction was cut along a plane through the centre of the core and the crystal axis. The shape of the p–n junction as revealed by copper-plating was determined with a microscope; the result is given in Fig. 7. This

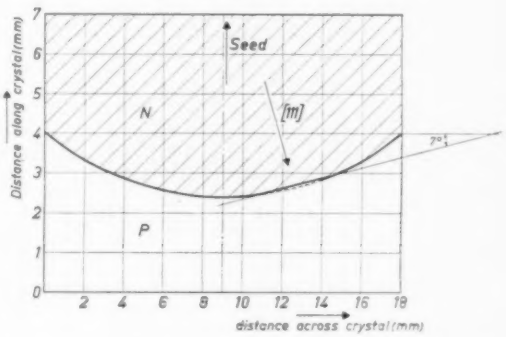


FIG. 7. Shape of a p–n junction in a crystal with $\theta = 7°$.

figure also shows a perfectly flat region on the curved growth interface. Additional measurements showed that the flat plane on the curved interface coincided indeed with a {111} plane. Thus it appears that the growth interface tends to stabilize in {111} planes. This can only occur where a ⟨111⟩ direction is approximately perpendicular to the interface. The place of the {111} facet corresponds exactly with the position of the resistivity core. One must therefore conclude that the origin of cores is closely connected with the formation of flat {111} facets on the growth interface. The variations in impurity concentration and consequently

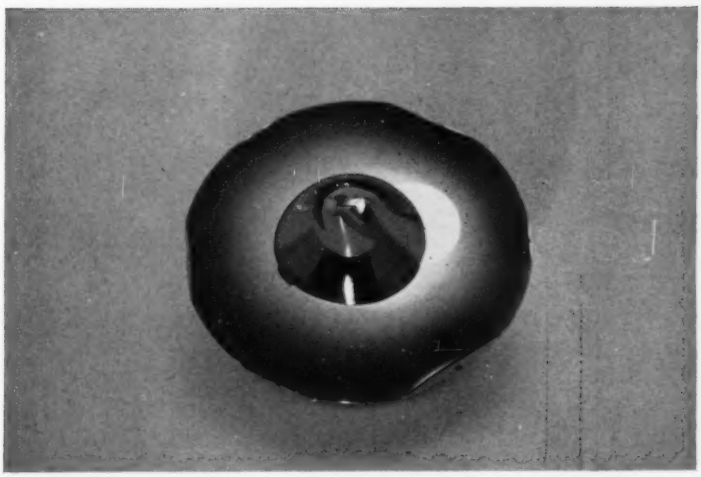


FIG. 6. Growth interface as revealed after rapid withdrawal of a crystal ($\theta = 7°$) from the melt.

[*facing p.* 206

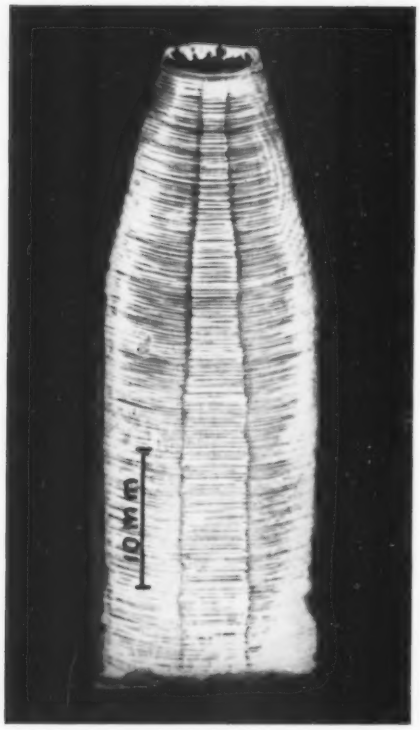


FIG. 8. Pulsed-plating pattern on a longitudinal section of a crystal with $\theta < 1°$.

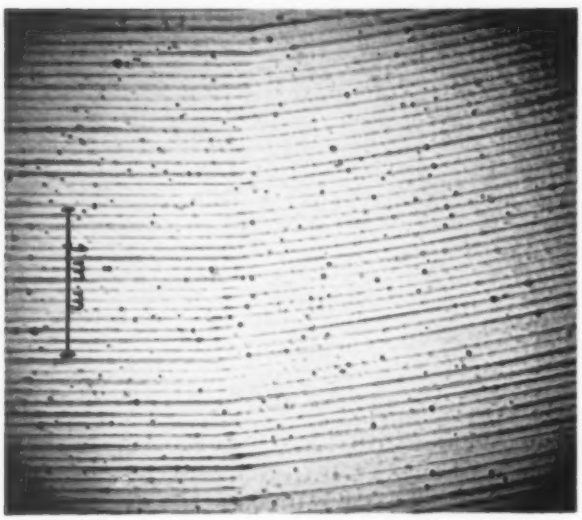


FIG. 10. Pulsed-plating pattern on a longitudinal section of a crystal with $\theta = 7°$ (rotated at 14 rev/min).

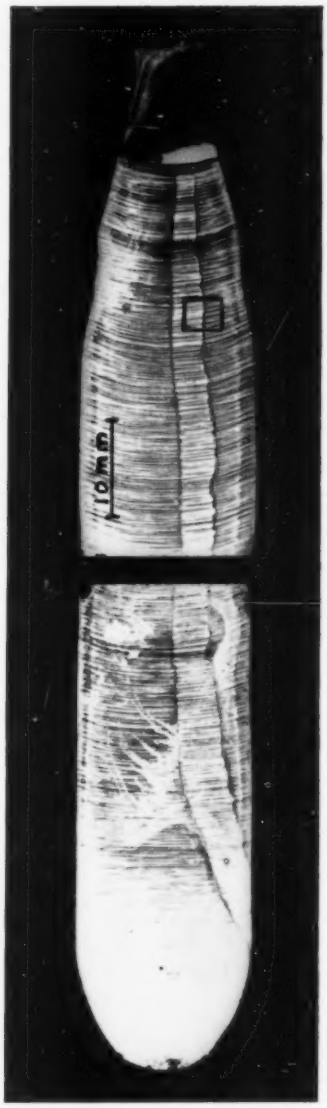


FIG. 9. Pulsed-plating pattern on a longitudinal section of a crystal with $\theta = 7°$.

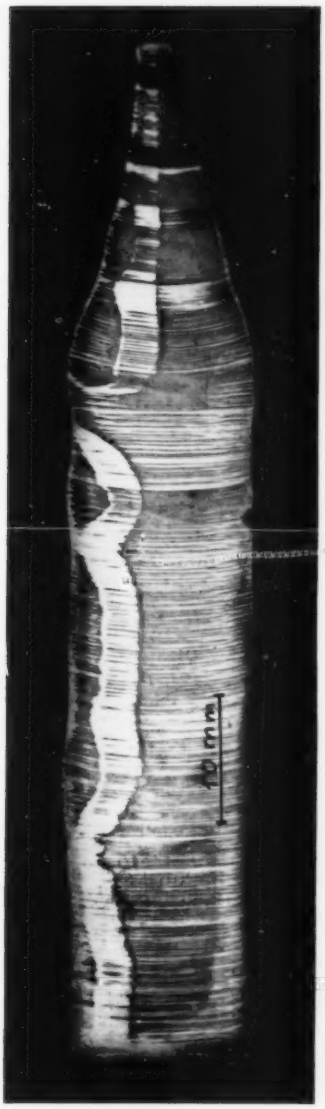


FIG. 11. Pulsed-plating pattern on a longitudinal section of a crystal ($\theta = 7°$) grown from a crucible as shown in Fig. 2(b).

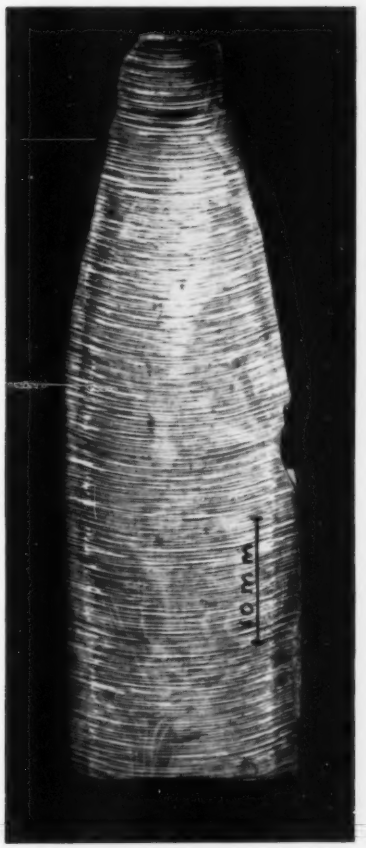


FIG. 12. Pulsed-plating pattern on a longitudinal section of a [100] crystal ($\theta = 55°$).

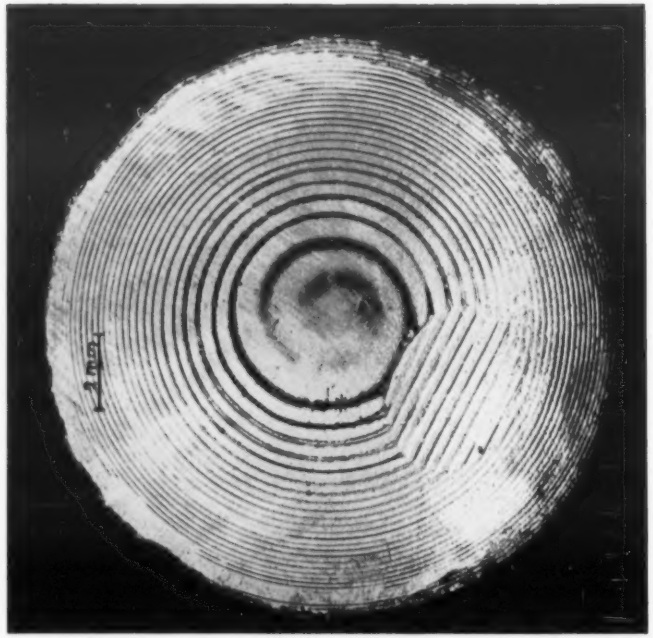


FIG. 13. Pulsed-plating pattern on a cross-section of a crystal with $\theta = 7°$.

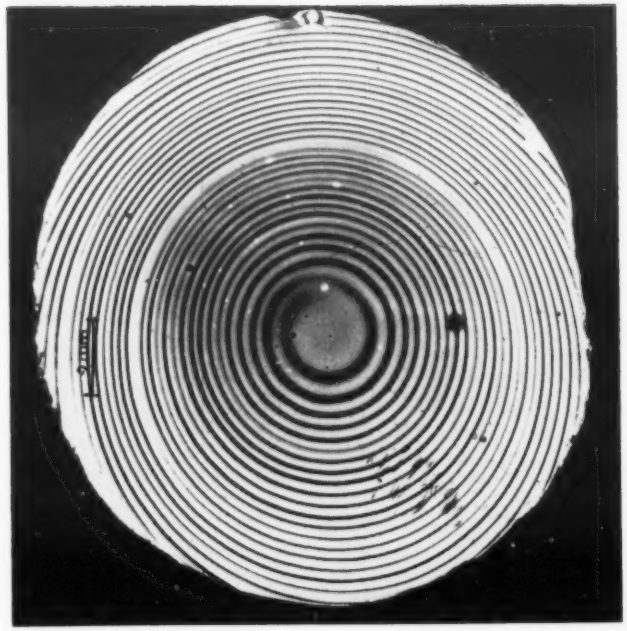


FIG. 14. Pulsed-plating pattern on a cross-section of a [100] crystal.

the resistivity variations are to be attributed to a difference in segregation constant at the flat part and at the curved part of the growth interface.

That the position of a core in a crystal depends on the growth direction and also on the shape of the growth interface is clearly demonstrated by the patterns obtained by pulsed copper-plating of longitudinal sections of some crystals. Fig. 8 shows the plating pattern of a crystal grown with $\theta<1°$. The pattern consists of a series of striations with the shape of the growth interface (see Section II). In the centre of the pattern a region, bounded by the vertical lines, can be seen in which the striations differ from that in the rest of the pattern. In the outer region the striations are curved and in the inner region, i.e. the core region, they are straight (see Fig. 10). This is in agreement with the foregoing discussions, because in this case the (111) plane is tangent to the growth interface in the centre of the crystal.

In a crystal with $\theta = 7°$ (see Fig. 9) the (111) plane coincides with the growth interface away from the centre of the crystal. A small part of the plating pattern of a similar crystal (as indicated by the little square in Fig. 9) is shown in Fig. 10 with a larger magnification. This figure shows clearly that at the core boundary the striations change from straight lines to curved lines. It appears from Fig. 9 that at the lower end of the crystal the growth interface gradually changes from convex to concave with respect to the melt, thus causing the core to grow out of the crystal. The same relation between the position of the core and the curvature of the interface is found in Fig. 11. This figure shows the plating pattern of a crystal grown with $\theta = 7°$ from a crucible as shown in Fig. 2(b). At those places in the crystal where the diameter decreased during growth the interface flattened, causing the core to move out of the crystal. The plating pattern (Fig. 12) of a [100] crystal ($\theta = 55°$) does not show any core, which is to be expected if the curvature of the interface does not coincide with a {111} plane at any point.

As was pointed out at the end of Section II, pulsed copper-plating reveals the intersection of the plated surface with a helical surface of different resistivity in the bulk of the crystal. The "steps" of this surface have the shape of the growth interface. When the latter was curved, the "steps" of the helical surface exhibit a similar curvature, so that a "dished" helical surface is obtained. If the pitch is small enough with respect to the curvature of the "steps" then the intersection of the surface with a plane *perpendicular* to its axis is a spiral. When part of the growth interface was flat, the corresponding part of the intersection spiral must be straight. The plating pattern (Fig. 13) on the cross-section of a crystal grown with $\theta = 7°$ does indeed show a spiral, the windings of which are locally straightened out to parallel straight lines, whereas the spiral on the cross-section of a [100] crystal (Fig. 14) does not show this feature. This confirms the fact that a [100] crystal, unlike a [111] crystal, does not have a flat facet on the growth interface.

c. *Influence of impurities*

Fig. 15 shows the result of a one-point resistivity measurement on a thin bar cut through the core

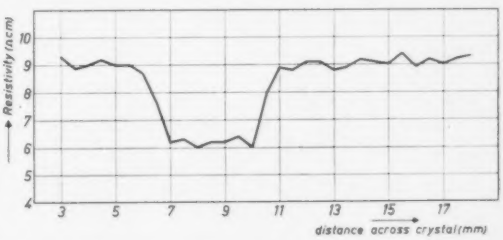


FIG. 15. Resistivity versus diameter of an antimony-doped [111] crystal with $\theta = 7°$ (one-point measurement).

parallel to the cross-section of an antimony-doped crystal. It may be seen that the resistivity changes abruptly at the core boundary while the resistivity in both regions is practically constant. The ratio of these resistivities ρ_p/ρ_c appears to be 1·45, where ρ_p and ρ_c are the resistivities outside and inside the core region. It was found that this ratio is independent of the doping level in the region from 0·05 Ω-cm to 20 Ω-cm for crystals grown with the same pulling speed. For other impurities a similar behaviour was found; however, the value of the ratio differs considerably for different impurities. This indicates that the segregation constant is different for the two regions. Consequently one may define a core-forming factor $\alpha = k_c/k_p$ in which k_c and k_p are the segregation constants inside and outside the core region. The core-forming factor α for several

impurities in germanium has been determined; the result is listed in Table 1; these values of α apply for a growth rate of 1 mm/min* and a rotation rate of 45 rev./min. It may be remarked that the core-forming factor of all impurities investigated is larger than unity, except for gallium.

Table 1. Core-forming factor, $\alpha = k_c/k_p$, of several impurities as observed in germanium crystals grown with a pull rate of 1 mm/min

Element	α
P	2·5
As	1·8
Sb	1·45
Bi	1·65
Ga	0·85
In	1·4
Tl	1·2

d. *Prevention of cores*

The data presented in this paper suggest two ways of preventing the formation of resistivity cores.

Firstly, one may grow a crystal in such a way that no {111} facets can be formed at part of the growth interface. This means that one has to choose the proper combination of the crystal orientation and the shape of the growth interface. It the interface is convex toward the melt, the angle θ between the growth direction and the [111] direction must be so large that a (111) plane nowhere coincides with the growth interface. If the interface is flat, as is, for instance, the case in the floating crucible described by GOORISSEN and KARSTENSEN[4], the orientation is less important. For if the growth direction coincides with the [111] direction the entire interface consists of a (111) facet, whereas a small angle θ will already be sufficient to prevent the formation of a (111) facet. In these two cases the resistivity is practically constant over the entire cross-section.

Secondly, resistivity cores may be prevented by doping with two or more impurities in such a way that the core formed by one of the impurities is compensated by the cores of the other impurities. In order to effect this the net donor or acceptor content must be the same in both regions, which requirement is fulfilled when $F = 1$ in the equation

$$F = \frac{\sum_i \alpha_{D_i} k_{D_i} C_{D_i} - \sum_j \alpha_{A_j} k_{A_j} C_{A_j}}{\sum_i k_{D_i} C_{D_i} - \sum_j k_{A_j} C_{A_j}} \qquad (1)$$

where C is the concentration in the melt, k the segregation coefficient and α the core-forming factor, while the subscript D denotes a donor impurity and A an acceptor impurity. To obtain compensation of the core with two donor or two acceptor impurities it is necessary that α is smaller than unity for one of the impurities and larger than unity for the other impurity. This method can consequently be used for p-type germanium by doping with gallium together with indium or thallium (see Table 1). Using gallium and indium in the melt with a concentration ratio of $C_{Ga}/C_{In} = 0{\cdot}03$, crystals were obtained with cross-sectional resistivity variations smaller than $\pm$ 3 per cent in spite of the occurrence of {111} facets on the curved growth interface as was revealed by pulsed copper-plating. It also follows from equation (1) that core compensation is possible by using an acceptor impurity together with a donor impurity with a different α.

For the sake of completeness it should be pointed out that if F in equation (1) is negative, the conductivity type of the core will be opposite to that of the remainder of the crystal. This may, for instance, be obtained by doping with phosphorus and gallium. The concentration ratio C_P/C_{Ga} required to obtain the desired value of F can be calculated from equation (1). As an example, the plating pattern of a p-type crystal with an n-type core is shown in Fig. 16.

IV. DISCUSSION

It appears from the experiments described here that the cross-sectional resistivity variations originate from the tendency of the solid–liquid interface of a growing germanium crystal to stabilize on {111} planes together with the dependence of the segregation constants of impurities upon crystal orientation. The occurrence of {111} facets on the growth interface may be attributed to the

* Preliminary experiments showed that α depends upon growth rate; for smaller growth rates, α is closer to unity. This effect however was not investigated in detail.

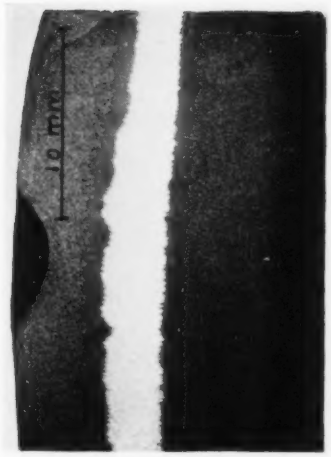


FIG. 16. Platinum-plating pattern of a *p*-type crystal with an *n*-type core.

[*facing p.* 208

dependence of the freezing rate upon crystal plane, as has been discussed by CHALMERS[15]. For germanium the [111] direction is the direction of slowest growth. Thus for growth at a (111) facet ("special surface") on the growth interface, more supercooling is required than at the remainder of the interface ("random surface"). It has been attempted to correlate this with the difference in segregation constant for both "surfaces". In Fig. 17 is shown part of the phase diagram of an impurity ($k_0<1$) in germanium. The slope $(dT/dC)_L$

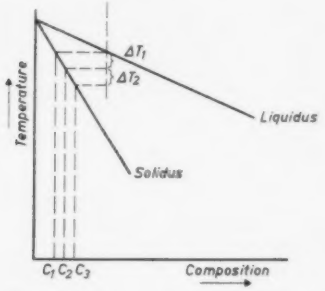


FIG. 17. Phase diagram of impurity ($k_0<1$) in germanium.

of the liquidus curve has about the same value for all impurities with an equilibrium distribution coefficient $k_0 \ll 1$; the slope of the solidus curve is related to the distribution coefficient by $(dT/dC)_S = (dT/dC)_L/k_0$. Let ΔT_1 be the supercooling required for growth on a "random surface" and $(\Delta T_1 + \Delta T_2)$ that for a "special surface". The impurity concentration in the freezing solid at the "random surface" will be between C_1 and C_2 and at the "special surface" between C_1 and C_3. The amount of impurity picked up by the solid would thus be larger at the flat part than at the curved part of the interface. This difference in concentration should further depend upon the slope of the solidus curve and consequently upon the distribution coefficient, the difference being larger for a larger k_0. Accordingly one would expect a larger core-forming factor α when k_0 is larger. In Fig. 18 the core forming factor α is plotted versus the distribution coefficient k_0 for the impurities investigated. It can be seen that, although most of the impurities behave in accordance with the above argument, bismuth and gallium deviate markedly. Gallium even behaves in an opposite manner to what one would expect. Because of these findings the above explanation cannot be the only one. There must be at least one other factor which accounts for the difference in the segregation constant for different crystal planes.

It has already been reported by HALL[16] that the segregation constant depends upon the crystal orientation when the growth rate is large enough.

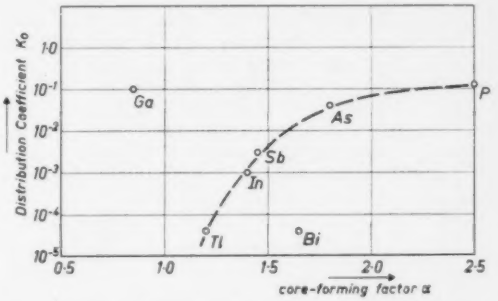


FIG. 18. Core-forming factor α versus distribution coefficient k_0 for several impurities in germanium.

For instance, in the case of antimony in germanium $k_{[111]} > k_{[100]} > k_{[110]}$; the differences being about 10 per cent at a growth rate of 1 mm/min. BRIDGERS and KOLB[17] found differences of the same order of magnitude. HALL[16] and BURTON *et al.*[2] found that the segregation constant also depends on growth rate. In order to explain this effect BURTON *et al.*[2] have given a theory based on the presence of a diffusion layer in front of the growth interface. The thickness δ of this layer is related to the segregation constant.

HALL suggested another possibility, namely that impurity atoms are preferentially adsorbed at the growth interface to a degree which varies with the crystal plane exposed to the melt. These "adsorbed" impurities are trapped in the solid at growth rates high enough to preclude their escape by solid-state diffusion. On the other hand BRIDGERS and KOLB[17] explain the variation of the segregation constant with crystal orientation by assuming that the thickness δ from BURTON'S theory depends upon crystal orientation.

From the present experiments it appears, firstly, that the variations of the segregation constant with orientation are much larger than those reported by HALL and by BRIDGERS and, secondly, that the segregation constant of some impurities (e.g. in-

dium) at a (111) facet is larger than that at the remainder of the growth interface, while for other impurities (e.g. gallium) the reverse is the case. This cannot be explained on the basis of the theory given by BURTON. The present results* support HALL's view that there is an "adsorption" of impurities at the growth interface, but the nature of this "adsorption" is not very clear.

Acknowledgements—The author wishes to thank Prof. D. POLDER, Mr. J. GOORISSEN, Dr. B. OKKERSE and Mr. P. PENNING for introducing him to this field and for helpful discussions. He is also much indebted to Dr. J. BLOEM for valuable advice concerning the pulsed-plating technique. His sincere thanks are due to Mr. A. HARTE and Miss C. J. M. R. Y. BAREL for their assistance with the experiments and to Mr. P. G. M. Vos and Mr. J. J. J. ZWEGERS for making the photographs.

REFERENCES

1. W. G. PFANN and K. M. OLSEN, *Phys. Rev.* **89,** 322 (1953).
2. J. A. BURTON, R. C. PRIM and W. P. SLICHTER, *J. Chem. Phys.* **21,** 1987 (1953); J. A. BURTON, E. D. KOLB, W. P. SLICHTER and J. D. STRUTHERS, *Ibid.* **21,** 1991 (1953).
3. W. F. LEVERTON, *J. Appl. Phys.* **29,** 1241 (1958).
4. J. GOORISSEN and F. KARSTENSEN, *Z. Metallk.* **50,** 46 (1959).
5. B. OKKERSE, *Handbook of Semiconductor Electronics*, 2nd Ed. (edited by L. P. HUNTER). McGraw-Hill, New York. To be published.
6. I. G. CRESSEL and J. A. POWELL, *Progr. Semiconductors* **2,** 137 (1957).
7. D. C. BENNETT and B. SAWYER, *Bell Syst. Tech. J.* **35,** 637 (1956).
8. A. J. GOSS and R. E. ADLINGTON, *Marconi Rev.* **22,** 18 (1959).
9. P. R. CAMP, *J. Appl. Phys.* **25,** 459 (1954).
10. B. OKKERSE. Private communication.
11. J. GOORISSEN. Private communication.
12. J. G. WILKES, *Proc. Instn. Elect. Engrs.* B**106,** 866 (1959).
13. G. K. TEAL and J. B. LITTLE, *Phys. Rev.* **78,** 647 (1950).
14. J. BLOEM. Unpublished results.
15. B. CHALMERS, *J. Metals, N.Y.* **6,** 519 (1954).
16. R. N. HALL, *Phys. Rev.* **88,** 139 (1952); *J. Phys. Chem.* **57,** 836 (1953).
17. H. E. BRIDGERS and E. D. KOLB, *J. Chem. Phys.* **25,** 648 (1956).
18. K. F. HULME and J. B. MULLIN, *Phil. Mag.* **4,** 1286 (1959).

* It should be mentioned that after completion of this paper HULME and MULLIN[18] reported the occurrence of {111} facets at the growth interface of InSb crystals. They also found a difference in the segregation constant of Te in InSb at the flat part and at the curved part of the interface.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 211–224. Printed in Great Britain

THE OPERATION OF JUNCTION TRANSISTORS AT HIGH CURRENTS AND IN SATURATION

C. A. MEAD

California Institute of Technology, Pasadena, California

(*Received* 2 *February* 1960; *in revised form* 19 *May* 1960)

Abstract—The lumped-model characterization for junction transistors has been extended to current ranges where the transistor operation may no longer be considered linear. A number of effects are considered which contribute to the nonlinear behavior. In all cases it is shown that the self-bias cutoff effect plays an important part in the deterioration of performance at high currents.

In alloy transistors, both normally and inversely biased, an exact solution for the form of the current gain as a function of collector current is given for the case where the injected density is small compared with the equilibrium base majority-carrier density. This solution is shown to be applicable to normally biased diffused-base units. A decrease in current gain is predicted due to the self-bias effect. The sources of lateral base current are bulk recombination and injection into the emitter, the relative amount being immaterial since both are proportional to the injected density. In the small-area alloy devices, it is seen that the injected density becomes large compared with the base majority-carrier density before the transistor performance is appreciably degraded. In such cases the lateral base current is seen to be normally dominated by non-unity emitter efficiency.

In all cases, surface recombination is assumed to be the dominant source of total base current, and the first-order correction to the linear theory is shown to exhibit a $1/(1+Ki_c)$ dependence. At higher currents all cases follow the observed $1/i_c$ proportionality.

The effect of nonlinear phenomena on saturation voltage is considered in some detail. The saturation voltage is shown to increase with drive current, again emphasizing the importance of avoiding large overdrive currents. The use of alloy transistors as low-level signal switches is discussed, indicating the necessity for high values of β and low values of K.

Résumé—La caractérisation des transistors à jonctions de modèle à constante localisée a été étendue pour des courants où le fonctionnement des transistors ne peut pas être considéré comme linéaire. On considère un certain nombre d'effets qui contribuent au comportement non linéaire. Dans tous les cas, on montre que la coupure d'autopolarisation joue un rôle important dans la diminution des performances à haut courant.

Dans des transistors alliés, aussi bien en fonctionnement normal qu'en polarisation, on donne une solution exacte pour la forme du gain de courant en fonction du courant du collecteur dans les cas où la densité d'injection est petite par rapport à la densité des porteurs majoritaires à l'équilibre. On montre que cette solution est applicable pour des unités ayant des bases diffusées normalement polarisées. On prévoit une diminution du gain du courant due à l'effet de l'autopolarisation. Les sources pour le courant des bases latérales sont la recombinaison interne et l'injection dans l'emetteur, la quantité relative est sans importance puisque toutes les deux sont proportionnelles à la densité injectée. Dans tous les dispositifs alliés à faible surface on trouve que la densité injectée devient grande par rapport à la densité des porteurs majoritaires de base avant que les performances des transistors soient diminuées d'une façon appréciable. Dans ce cas le courant de base latérale se trouve normalement dominé par l'efficacité de l'émetteur différent de l'unité.

Dans tous les cas, la recombinaison en surface est supposée être la source prédominante du courant de base total et la correction de premier ordre à la théorie linéaire montre une dépendance en $1/(1+Ki_c)$. A courant plus élevé tous les cas suivent la relation observeé $1/i_c$.

L'effet des phénomènes non linéaires pour le voltage de saturation est considéré en détails. On montre que le voltage de saturation augmente avec le courant d'entrainement. On discute l'utilisation des transistors alliés pour des interrupteurs à bas niveaux de signal en indiquant la nécessité d'avoir des valeurs de β élevées et de faibles valeurs de K.

Zusammenfassung—Die Darstellung von Flächentransistoren durch Ersatzschaltbilder vom Typ des "lumped model" wurde auf Strombereiche ausgedehnt, in denen der Transistor nicht mehr linear arbeitet. Mehrere Effekte werden betrachtet, die zu dem nicht-linearen Verhalten beitragen.

Es zeigt sich in allen Fällen, dass die zwangsweise Vorspannung des Emitters infolge des Spannungsabfalles in der Basis eine wichtige Rolle in der Verschlechterung der Leistungsfähigkeit des Transistors bei hohen Strömen spielt.

Für Legierungstransistoren, sowohl normal als auch umgekehrt vorgespannte, wird eine exakte Lösung für die Stromverstärkung als Funktion des Kollektorstromes angegeben und zwar für den Fall, wenn die Dichte der injizierten Ladungsträger klein gegenüber der Gleichgewichtsdichte der Majoritätsträger in der Basis ist. Es wird gezeigt, dass die Lösung auch für normal vorgespannte Transistoren mit diffundierter Basis anwendbar ist. Eine Abnahme der Stromverstärkung auf Grund der zwangsläufigen Vorspannung des Emitters wird voransgesagt. Die Ursachen des nach den Seiten abfliessenden Basisstromes sind Volumenrekombination und Injektion in den Emitter. Der relative Beitrag der beiden Effekte ist an sich unwesentlich, da beide proportional der Dichte der injizierten Ladungsträger sind. Es zeigt sich, dass in den Legierungstransistoren mit kleiner Fläche die Dichte der injizierten Ladungsträger grösser wird im Vergleich zu der Dichte der Majoritätsträger in der Basis, ehe die Leistungsfähigkeit des Transistors merklich nachlässt. In solchen Fällen erweist es sich, dass normalweise der sich von den Einheit unterscheidende Emitterwirkungsgrad den Basisstrom beherrscht.

In allen Fällen wird Oberflächenrekombination als vorherrschende Ursache des totalen Basisstromes angenommen, und es wird gezeigt, dass die Korrektur der linearen Theorie in erster Näherung eine Abhängigkeit proportional zu $1/(1+Ki_c)$ gibt. Bie höheren Strömen wird in allen Fällen die beobachtete Proportionalität zu $1/i_c$ befolgt.

Der Einfluss nicht-linearer Erscheinungen auf die Sättigungsspannung wird im einzelnen betrachtet. Es wird gezeigt, dass die Sättigungsspannung mit dem Steuerstrom wächst, woraus man wiederum ersieht, wie wichtig es ist, stark überhöhte Steuerströme zu vermeiden. Die Anwendung legierter Transistoren als Schalter für kleine Leistungen wird diskutiert, wobei sich die Notwendigkeit hoher β-Werte und niedriger K-Werte erweist.

INTRODUCTION

THE analysis of junction-transistor switching performance has required the development of characterization techniques considerably more advanced than those used for small-signal analysis. Perhaps the most satisfactory technique which has emerged is that of the linear lumped model.[1, 2]

It has been pointed out that the lumped-model approach provides a deep understanding of the physical principles underlying transistor operation. One is naturally led to ask if such an approach would also provide a similar understanding of the nonlinear behavior of junction transistors which predominates as current densities are pushed higher and higher. In the sections which follow, the nonlinear theory will be investigated in some detail, and modifications of the linear theory which are necessary under a number of conditions encountered in practice will be established. Throughout the discussion a p–n–p transistor is assumed in the interest of clarity and simplicity of notation.

The lumped model for a p–n–p alloy transistor with very high-conductivity collector and emitter regions is shown in Fig. 1. The symbols have the same significance as passive electrical-circuit elements, except that they relate the excess minority-carrier density ρ, rather than voltage, to the current. In this model ρ_1 represents the excess density near the emitter, given by

$$\rho_1 = p_n[\exp(qv_{eb}/kT)-1]$$

where v_{eb} is the emitter-base junction voltage; C_1

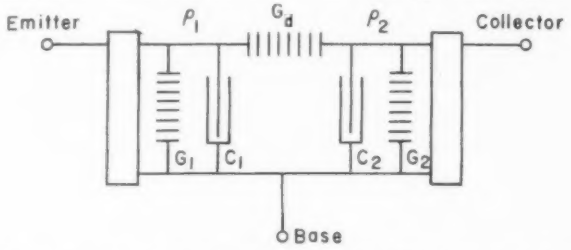


FIG. 1. Complete two-section lumped model of junction transistor with high-conductivity collector and emitter regions.

and G_1 represent storage and recombination near the emitter, G_d represents diffusion from emitter to collector, ρ_2 is the excess density near the collector, given by

$$\rho_2 = p_n[\exp(qv_{cb}/kT)-1]$$

where v_{cb} is the collector-base junction voltage; C_2 and G_2 represent storage and recombination near the collector. Both voltages are taken positive when the junction is forward biased. A very de-detailed discussion of the linear lumped model is given in Ref. 2.

SUMMARY OF NONLINEAR EFFECTS

Let us now investigate, in general, the effects which would be expected to contribute to the nonlinear behavior of a junction transistor. Although a complete list would necessarily be of much greater length than utility, the following effects are of some importance in present devices and merit careful consideration.

Depletion layer recombination

At low injection levels in silicon devices, junction operation is dominated by carrier recombination within the depletion layer. In germanium devices and in silicon devices operating at higher levels, this effect is negligible.[3–5] Since we are specifically interested in the high-level case, we shall assume that the effects of depletion-layer recombination are small. However, some discrepancies will be found to exist between the high-level theoretical expressions and data taken at relatively low level, which may be explained in terms of this omission, but which are not important for the purposes of this discussion. The effect of such recombination on the lumped model is to introduce conductances in parallel with G_1 and G_2 which decrease rapidly as ρ_1 and ρ_2 become appreciable.

Drift-enhanced diffusion flow

When minority carriers are injected into a region ot semiconductor, the condition of near-charge neutrality requires an equal number of additional majority carriers. A gradient in excess minority-carrier density therefore implies a gradient in majority-carrier density. If the majority-carrier current is constrained by means of junctions, the diffusion-current tendency implied by this density gradient must be canceled by an equal and opposite drift-current tendency; hence an electric field is developed. As long as the injected density is small compared to the equilibrium majority density, the drift component of minority carrier resulting from this electric field is small compared to the diffusion component. However, at high injection levels, the drift component becomes appreciable, resulting in an effective diffusion constant larger than at low densities. At very high levels, the effective diffusion constant approaches twice its low-level value.[6] As we shall later separate the high- and low-level cases, it will not be necessary to consider the transition in detail. However, since the low-level limiting form of the solution is of interest, it must be borne in mind that the high-level value of G_d will be twice the low-level value due to this effect.

Recombination rate

For most purposes the rate of recombination of minority carriers may be considered proportional to the product of majority- and minority-carrier densities. However, in semiconductor material as commonly used, recombination occurs principally at traps or recombination centers, and the recombination statistics are altered.[4, 5, 7] This effect is chiefly responsible for the rapid decrease in current gain at low current, but may be neglected at higher current levels.

Conductivity modulation

The conductivity of a semiconductor is usually taken as proportional to the equilibrium density of majority carriers which, under normal conditions, is equal to the density of donor or acceptor atoms. However, at high levels the density of majority carriers is increased as the density of minority carriers becomes appreciable. Hence the number of carriers available for current flow is increased and the conductivity of the material is increased. Ohmic resistance effects (for example, the base resistance and series collector resistance of a transistor) are altered to the extent that these effects arise in the active region of the transistor. The conductivity σ of the material becomes

$$\sigma = \sigma_0\left[1+(1+b)\frac{\rho}{n_n}\right] \qquad (1)$$

where σ_0 is the low-level conductivity and b is the ratio of minority- to majority-carrier mobilities.[8]

Junction injection efficiencies

When a p–n junction is forward biased, in addition to the holes injected from the p-region into the n-region, electrons are injected from the n-region into the p-region. The ratio of the in-injected densities in the two regions is inversely proportional to the ratio of the majority-carrier densities. If the equilibrium majority density of one region is much greater than that of the other, the injected density will become comparable to the majority density at much lower levels in the low-conductivity region. Hence the ratio of injected densities will vary with the injection level. This effect becomes significant, for example, at a forward-biased junction of a transistor.

Although the contribution to the total base current in a junction transistor due to non-unity injection efficiency is normally small compared with that due to surface recombination, the lateral base current within the active area of the device is often principally due to injection effects and hence of great importance in connection with the self-bias cutoff effect.

As previously noted, in many transistor junctions the conductivity of the transistor base region is relatively low compared with that of the other region in question. In this case the current density J injected from the base region into the other region may be shown to be related to the excess density ρ in the base region as follows:[3, 9]

$$J \sim \rho\left(1+\frac{\rho}{n_n}\right) = p_n[\exp(qv/kT)-1] \qquad (2)$$

where

p_n is the equilibrium minority-carrier density in the base, k is Boltzmann's constant, T is the absolute temperature and v is the voltage across the junction, being taken as positive when the junction is forward biased.

Density-crowding or self-bias effect

Because of volume recombination and non-unity junction efficiencies, majority-carrier current is caused to flow laterally in the base region. This current must flow through the sheet resistance of the base region and hence causes a lateral potential gradient in the base region. This gradient is in such a direction as to decrease the junction forward bias in the center of the base region. Hence at high levels the injected carriers are essentially confined to the outer edge of the base region.[10–12] The major consequence of this effect is a reduction in the effective area of the active region of the transistor. In many transistors the high-level behavior is dominated by this effect. In this paper we will consider the results of density crowding in considerable detail.

Thermal effects

When a junction transistor is operated at high current densities, heat is evolved at junctions where the applied reverse bias is appreciable, and also in the ohmic resistances associated with the various regions. Changes in temperature of the device caused by heat generated in this manner cause changes in the electrical properties of the unit which are often very difficult to distinguish from the high-level phenomena previously discussed. Junctions in a transistor or diode are often assigned a certain thermal time constant and thermal resistance to the mounting base of the device, and it is thus implied that the junction itself possesses a certain thermal capacitance, which may be fairly large. However, the flow of heat in a solid is basically a diffusion process and it should be recognized that the lumped approximation implied by the thermal time constant and thermal resistance is valid only under conditions where the thermal gradients are not too severe. Under conditions of very high power input for very short periods, such as during pulse measurements, the lumped approximation would predict a very small change in temperature. In reality, since the power is dissipated only within the junction, and hence is localized to a very small volume, the actual local temperature may rise to a very high value. Such a temperature increase causes violent changes in the equilibrium minority-carrier density as well as less drastic changes in lifetime, mobility and surface-recombination properties. Although we shall not treat these effects in detail, we must always be aware that experimental results may be colored by their presence.

Hole–electron scattering

As the injected carrier density reaches very high levels, a decrease in the mobility of minority carriers is to be expected due to hole–electron scattering. This effect is treated in considerable detail in a paper by FLETCHER[13], where the

mobility of holes and electrons is plotted as a function of minority-carrier excess density. Although these results have not been explicitly incorporated into the present work, it should be noted that since the lumped-model approach has been used throughout, the determination of injected densities is a necessary step in the solution of any high-level problem. Like FLETCHER[10] and others,[11, 12] we have assumed the mobility to be constant. However, a good estimate of the effects of hole–electron scattering may be obtained using the computed values of injected density in connection with FLETCHER's data.

Outline of analysis

In order to gain a basic understanding of transistor high-current operation we have briefly considered a number of the more important effects which contribute to its nonlinear behavior. We shall now restrict our attention to a number of specific cases of practical interest.

Of the bewildering array of transistors now available, we shall consider only the alloy transistor, although many of the results are applicable to other types as well. Due to the self-bias effect we expect nonlinear performance even at injection levels low compared with the majority-carrier density. For many alloy units, much of the operation is performed under very high-level conditions. The corresponding analysis must include the effects of decreased injection efficiency and base conductivity modulation, which considerably modify the form of the self-bias effect, and are treated in some detail.

Finally, operation in the saturation region is considered, where the injection efficiency of both emitter and collector is seen to be important. The electric field across the base region due to majority-carrier density gradient may also make important contributions to the saturation voltage at high levels.

NONLINEAR OPERATION: NORMAL BIAS

The basic equations describing the operation of a junction transistor under conditions where carrier density crowding must be taken into account will now be derived. The analysis is presented for a normally connected transistor but is equally valid for the inverse connection. A typical transistor geometry is shown in Fig. 2. The structure is assumed two-dimensional, i.e. independent of the co-ordinate normal to the page. Thus, the analysis will be accurate for stripe and ring configurations and interleaved-finger power transistors. It will also be approximately correct for dot configurations at high levels where the current is confined principally to the edge of the emitter. Initially we shall solve for the distribution of minority-carrier density in the base region caused by the self-bias effect as a function of the distance from the edge of the emitter x. This information will allow us to calculate the base and collector current, and hence the behavior of current-transfer ratio as a function of current level. Another way of expressing the solution is in terms of the variation of the lumped-model parameters as a function of the excess density.

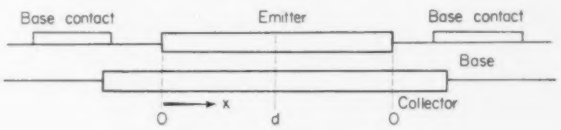


FIG. 2. Structure of alloy transistor.

Low-level approximation

As noted before, if $\rho \ll n_n$, both the electron current injected into the emitter and the bulk recombination may be considered proportional to the excess density. Hence the lateral (x-directed) base current flowing outward from the interior base region may be written as:

$$i(x) = A \int_d^x \rho(x')dx' \tag{3}$$

where A is the sum of two terms, the first due to volume recombination, and the second due to non-unity emitter efficiency.

The lateral base current flowing through the

sheet resistance of the base creates a lateral voltage drop within the base region. This self-bias voltage drop is in such a direction as to decrease the forward-bias voltage applied to interior portions of the emitter-base diode.[10] The voltage $v(x)$ between the base at the outer edge of the emitter ($x = 0$) and at some interior point x may be written as:

$$-v(x) = R \int_0^x i(x')dx' \tag{4}$$

where R is the sheet resistance of the base. (Note that A and R may be be considered constant only when $\rho \ll n_n$; conductivity modulation of the base region and the decrease in emitter efficiency create important variations which will be treated later.)

The effect of the self-bias voltage $v(x)$ is to decrease the excess hole density injected into the base region from its value at $x = 0$. As we have restricted our attention to the case $p_n \ll \rho \ll n_n$, we may write the injected density in terms of the self-bias voltage $v(x)$ from equation (2)

$$\rho(x) = \rho(0) \exp[-qv(x)/kT] \tag{5}$$

Both the mobility and diffusion constant of minority carriers have been assumed independent of excess density, as noted earlier. Hence each volume element may be considered as a separate transistor, and the total diffusion current through each element is proportional to the density gradient, or in the special case where the collector is reverse biased, to the emitter excess density at the element in question. Integrating the contributions to the diffusion over all elements, the total collector diffusion current per unit length of the structure i_c may be written

$$i_c = F \int_0^d \rho(x')dx' \tag{6}$$

where F is the carrier diffusion constant divided by the base width.

The formulation of the basic equations necessary for determining the transistor high-level operating characteristics is now complete. These equations with appropriate modifications will also be used for cases treated in later sections.

It may be verified by direct substitution that the solution of this set of simultaneous, nonlinear, integral equations (3–5) is[11]

$$i(x) = B \tan \frac{(d-x)}{s} \tag{7a}$$

$$\rho(x) = \rho(d) \sec^2 \frac{(d-x)}{s} \tag{7b}$$

where

$$B^2 = \frac{2kT\,\rho(d)A}{qR} \text{ and } s^2 = \frac{2kT}{\rho(d)\,ARq}$$

Since the lateral base current evaluated at the edge of the emitter $i(0)$ is proportional to the area under the density curve, and since the total collector diffusion current (i_c) is also proportional to the area under the density curve, the ratio of the two is a constant,

$$i_c = \frac{F}{A} i(0) \tag{8}$$

Thus one might expect the forward-current gain β of the transistor to be constant. However, it should be recalled that the effects of bulk recombination and emitter efficiency on β are normally not appreciable compared with the effect of surface recombination.[9] Even an approximate analysis of the surface-recombination current is extremely involved and becomes so entangled with the exact geometry of the specific transistor as to be useless in general. However, it may be shown that, for a wide variety of practical cases, the surface recombination may be considered nearly proportional to the excess density at the edge of the emitter.[14] This amounts to assuming G_1 independent of ρ. Hence, if for the purpose of determining the total base current i_b we neglect bulk recombination and the effects of non-unity emitter efficiency, i_b is proportional to ρ_1. We shall now investigate the dependence of G_d on the injection level. From equations (7) we may write

$$\rho(d) = \rho_1 - \frac{3}{2} \frac{K}{G_{d_0}} i_c^2 \tag{9a}$$

where

$$K = \frac{qRAd}{3kTF} \tag{9b}$$

Let us now investigate the special limiting case where $i_c^2 K$ is very small compared with $\rho(d)G_{d_0}$ and the excess density is uniform across the base

region. Under these conditions d/s is small and we may approximate the tangent by its argument

$$i_c \simeq \frac{F}{A}\frac{B}{s}d = Fd\rho(d) \simeq Fd\rho(0)$$

Hence the value of G_d at low levels (G_{d_0}) is seen to be

$$G_{d_0} = Fd$$

Returning to the more general case, equations (7) and (9) may be written

$$i_c = \sqrt{\left(\frac{2\rho_1 G_{d_0}}{3K} - i_c^2\right)} \tan \frac{3K}{2} \sqrt{\left(\frac{2\rho_1 G_{d_0}}{3K} - i_c^2\right)} \quad (10)$$

In order to generalize the result of this analysis, we shall express the quantities of interest in dimensionless normalized form

$$\rho^* = \frac{3}{2}\rho G_{d_0} K$$

$$i_c^* = \frac{3}{2} i_c K$$

$$G_d^* = \frac{G_d}{G_{d_0}}$$

Using this notation, equation (10) becomes

$$i_c^* = \sqrt{(\rho_1^* - i_c^{*2})} \tan \sqrt{(\rho_1^* - i_c^{*2})} \quad (11)$$

The radical may be considered as a parameter by which the values of ρ_1^* and i_c^* are connected. It is also a significant measure of the extent to which the carriers are crowded to the edge of the emitter. The normalized values, ρ^* and i_c^*, of the density and collector current may now be plotted and are related as shown in Fig. 3.

An important implication of these results is the ability to present a unique form for the dependence of i_c on ρ_1 with only one normalizing constant to be determined. As in the linear case, we find the combination ρG_{d_0} a convenient one with which to deal.[2]

The relation between i_c^* and ρ^* is seen to be linear at values of i_c^* much less than unity. At higher levels i_c^* increases only as the square root of ρ^*. The variation of the effective diffusion conductance with excess density and also with collector current is shown in Fig. 4. Again only the

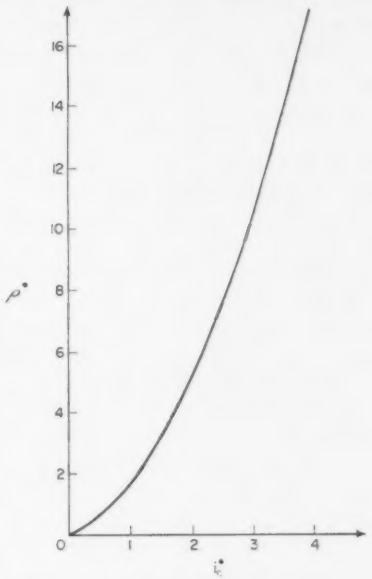


FIG. 3. Dependence of collector current upon excess density.

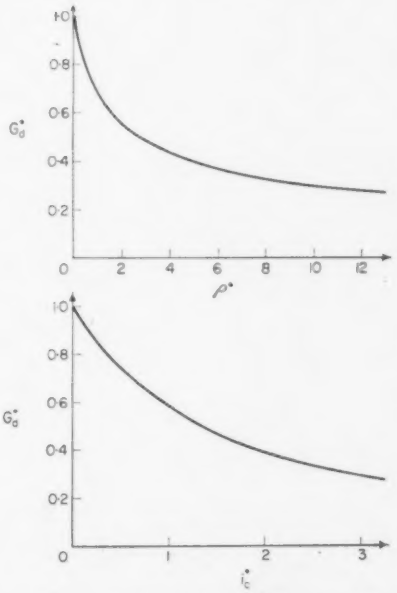


FIG. 4. Theoretical variation of effective diffusion conductance.

constant K (in addition to the linear lumped-model element values) is needed to characterize the transistor completely. The inverse of this constant, given in units of current, is a valuable figure of merit for transistors with which the present approximation is valid. From a circuit standpoint this value represents the current at which the transistor has begun to fail, and therefore provides a reliable comparison of the current-handling capacity of various units.

From a device-design point of view, we are interested in the requirements on the physical parameters of a transistor to make the figure of merit as large as possible. Since the figure of merit is inversely proportional to K we should strive for large values of F and small values of R, A and d.

The decrease in G_d at higher levels is due to a decrease in the effective area of the transistor due to the density crowding or self-bias effect. A very large reduction in G_d may be observed while the injected density is still small compared with the equilibrium majority density, n_n. In large-area power transistors this condition is very prevalent.

Experimental data taken on an alloy power transistor of essentially two-dimensional construction are shown in Fig. 5, together with the theoretical relation as given in Fig. 3. The value of K has been determined to give the best fit. It can be seen that the results agree within experimental error over a very wide current range.

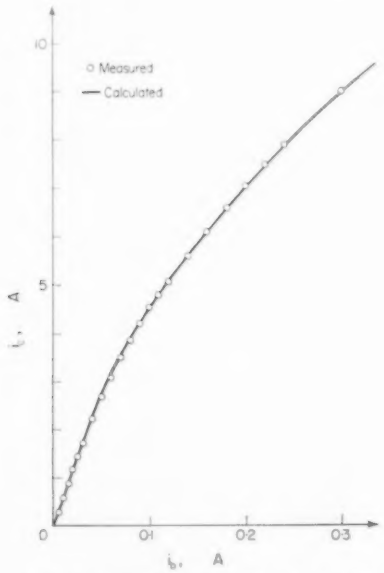


FIG. 5. Experimental verification of Fig. 3.

(1) *Near-uniform approximation.* Closed-form solutions for the relation between i_c and ρ_1 may be obtained for the two limiting cases of very slight and very severe crowding. For the slightly crowded case the diffusion conductance may be written

$$G_d \simeq \frac{G_{d_0}}{1+K\rho_1 G_{d_0}} \tag{12}$$

Since the condition for validity of this approximation is that the correction term involving K be small compared to unity, we may obtain an approximate expression for the current gain by substituting

$$i_c \simeq \rho_1 G_{d_0}$$

in the correction term. Again the value of G may be assumed constant and the currrent gain may be written

$$\beta \approx \frac{\beta_0}{1+Ki_c} \simeq \beta_0(1-Ki_c) \tag{13}$$

This expression is very significant since it gives the form of the falloff of β in the most useful range of operation.

(2) *Severely crowded approximation.* Under conditions of severe crowding, the injected density and collector current are related by the following expression:

$$\rho_1^* \simeq i_c^* + \frac{\pi^2}{4} \qquad \rho^* \gg 1$$

In the range where this approximation is valid, the constant term may be neglected without appreciably affecting the accuracy. Hence the diffusion conductance may be written:

$$G_d^* \simeq \frac{1}{i_c^*}$$

or

$$G_d \simeq \frac{2G_{d_0}}{3Ki_c} \tag{14a}$$

which implies:

$$\beta \simeq \frac{2\beta_0}{3Ki_c} \tag{14b}$$

It is interesting that this expression agrees with the high-current limit of equation (13) within the

factor of 3/2, and the conclusions to be drawn from either expression are the same.

A plot of $\rho^*(0)$ as a function of i_c^* is shown in Fig. 6, together with the asymptotic forms at low and high currents. It may be seen that results

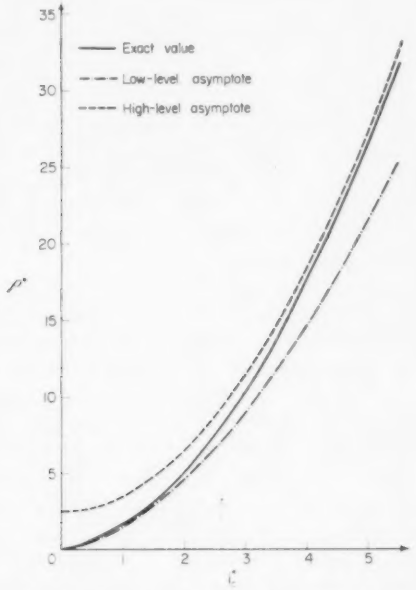


FIG. 6. Asymptotes to the relation of Fig. 3, for slight and severe crowding.

obtained by means of the asymptote are never in error more than about 10 per cent, the worst case occurring when $\rho^* \approx 10$ and convergence is quite rapid at lower or higher values.

High-level approximation

The foregoing expression for the figure of merit indicates that the self-bias cutoff effect or emitter density crowding, as discussed, would be most noticeable in transistors with large emitter dimensions, namely, power transistors. Indeed, the effect was first treated in detail by FLETCHER[10] in connection with power transistors. However, his analysis assumed infinite emitter width and hence did not treat the transition from uniform flow to the crowded condition. It is clear from equations (9) and (13) that for a given emitter area the current-carrying capability varies inversely with the emitter width. Hence the most efficient design for such units is an interleaved system of narrow base and emitter stripes or rings.[8, 10, 15] As the emitter width is decreased, the self-bias effect becomes less noticeable and current densities are pushed higher and higher. In many transistors now available, the assumption of $\rho \ll n_n$ is very poor; in fact the other extreme $\rho \gg n_n$ is often valid over the greater part of the range of operation. Thus, we must re-examine the analysis including the effects of large injected density.

At high injection levels two major changes in the physical operation occur:

(1) The emitter efficiency decreases as shown in equation (2). This decrease may be quite large, and in general at high levels the current injected into the emitter region will be the dominant source of lateral base current inside the area covered by the emitter.

(2) The resistivity of the base region decreases due to conductivity modulation caused by the injected minority carriers, as shown in equation (1). Hence R may not be considered a constant, and the variation with ρ must be shown in the integral.

If we assume that all the lateral base current is due to the decrease in emitter injection efficiency and if we neglect bulk recombinations, we may directly apply the expression for current density given in equation (2). The expression for lateral base current then becomes

$$i(x) = A \int_d^x \rho(x') \left(1 + \frac{\rho(x')}{n_n}\right) dx' \qquad (15)$$

The lateral voltage drop caused by this current is also affected by the conductivity modulation of the base region. Since the effective resistivity of each element of base region is reduced as given by equation (1) we may write

$$-v(x) = R \int_0^x \frac{i(x')}{1 + \dfrac{(1+b)\rho(x')}{n_n}} dx' \qquad (16)$$

The junction injection law is also modified at high levels in accordance with equation (2). We may assume as before $\rho \gg p_n$

$$\rho(x)\left[1 + \frac{\rho(x)}{n_n}\right] = \rho(0)\left[1 + \frac{\rho(0)}{n_n}\right] \exp[-qv(x)/kT] \qquad (17)$$

Further analysis becomes very difficult, unless another simplification is achieved. Since we are primarily interested in the high current or high injection-level limit on the devices, and since ρ becomes equal to n_n at quite low current levels in many devices,[4, 5] we shall make the assumption of high injection level ($\rho \gg n_n$), realizing that the results will break down at the lower levels, where the previous analysis may be applied. As is often the case with engineering approximations such as these, the regions of mathematical validity for the two approximations are widely separated. However, in practice, the approximations are reasonably good into the intervening range and quite a good representation of the overall performance may be made by the use of these relatively simple cases. Equations (15–17) may be converted to the single differential equation

$$\frac{d^2\rho(x)}{dx^2} = \frac{qRA}{(1+b)kT}\rho(x)^2 \tag{18}$$

Two approximate solutions of this equation will be considered.

(1) *Near-uniform approximation.* The first solution is derived for conditions where the crowding is not severe. In this case the diffusion conductance may be written[2]

$$G_d = \frac{G_{d_0}}{1+K\rho_1 G_{d_0}} \tag{19}$$

where in this case the current falloff factor K becomes

$$K = \frac{qRAd}{3(1+b)kTF}$$

which is seen to differ from the low-level case of equation (12) only by the constant $(1+b)$. The fact that a unique form is available for the two cases is quite significant for the circuit engineer. It permits the nonlinear characterization of the transistor to first order without knowledge of whether the unit is operating at high or low levels, simply by a measurement of the constant K.

(2) *Severe-crowding approximation.* In the case of transistors with wide emitters, severe crowding already exists as ρ becomes comparable with n_n. If we may assume severe crowding, an exact solution to equation (18) is possible.

$$i_c = 2F\sqrt{\left(\frac{6(1+b)kT}{qAR}\rho_1\right)}$$

which indicates a value of G_d which is inversely proportional to the collector current. The base current may be written in the form

$$i_b = \rho_1\left[G_s + \frac{A}{5n_n}\sqrt{\left(\frac{qRA}{6(1+b)kT}\right)}\rho_1^{3/2}\right] \tag{20}$$

where G_s is the effective surface recombination conductance. As noted before, in most cases the surface-recombination term will dominate until very high levels are reached. Under these conditions, β is again proportional to $1/i_c$ as in the other cases.

It should be pointed out that, in many transistors where the base resistivity is quite high and the emitter is very heavily doped, the approximations of $\rho \gg n_n$ in the base and low injection level in the emitter are valid over a very wide current range, and hence the analysis assumes a greater importance. However, in other cases where the ratio of the conductivities in the two regions is not nearly so large, the operation may vary smoothly from the low-level case to the condition of high injection in both regions.

It seems unnecessary to deal with the correlation between emitter efficiency and current-gain falloff as widely treated in the early literature. However, because of the widespread misconception of the principles involved, perhaps a word here is in order. From the results of the foregoing analysis it is clear that the $1/i_c$ dependence of β which was observed is to be expected as a result of the self-bias cutoff effect when the base current is completely dominated by surface recombination. Since the surface recombination is known to dominate in a great many transistors, especially earlier units, it is inconceivable that the base current due to decreased emitted efficiency could have been responsible for the rapid falloff in β which was observed. The great improvement which was effected by the use of highly doped emitter materials was due to the decrease in lateral base current and hence a decrease in the self-bias cutoff effect which resulted in a more uniform distribution of carriers across the emitter surface. This improvement resulted in a marked increase in collector current for a given value of injected

density. In other words, the observed variation in β was primarily the result of a decrease in G_d due to emitter crowding, not an increase in G_1. The necessity for an extremely large value of emitter efficiency is now apparent, since the small internal lateral base current is much more important than previously recognized.

As mentioned before, the results of this section are directly applicable to alloy transistors operating in either the normal or inverse connection, the low-level expressions are also applicable to normally biased diffused-base units. However, it should be noted that many alloy structures utilize a different doping material for the collector junction than for the emitter, and therefore the constant A will not, in general, be the same for the two connections. Also, the collector is normally considerably larger than the emitter, making the surface-recombination conductance G_2 larger than G_1.

SATURATED CONDITIONS; ALLOY TRANSISTOR*

One of the most outstanding and useful properties of a junction transistor is its unique ability to saturate. Under saturated conditions the minority-carrier density at the collector builds up to the point where the gradient of minority carriers is just sufficient to supply the necessary collector current. As the density curve rises higher, the saturation voltage becomes lower. The limit is reached when the total recombination in the base and injection into the emitter and collector become equal to the base driving current. In the linear theory the exact source of the base current was not important. However, as high-level conditions are reached, the self-bias effect crowds the carriers toward the outer edge of the collector and hence a large part of the collector junction may be essentially cut off. A rapid increase in base current for a given degree of saturation is caused by the large exposed collector area which has no facing emitter junction and hence carriers are injected, only to be lost in the base region.[17]

With the alloying techniques in common use, the base region outside the emitter is considerably wider than inside the active area of the device, and crowding effects may be neglected in this region. Suppose both emitter and collector junctions are forward biased such that at $x = 0$ their respective injected densities are ρ_1 and ρ_2. The analysis of this case is much the same as that used in connection with the normally biased transistor. Under the assumption that the value of A is the same at emitter and collector, the results obtained indicate a behavior of G_d which is identical with that given in the normal biased case, provided ρ is replaced by $\rho_1+\rho_2$. As before, two special cases are considered: the near-uniform and severely crowded approximations. In the former case a closed-form solution for the saturation voltage as a function of i_c and i_b may be obtained.

$$v_{sat} \simeq \frac{kT}{q} \ln \left[1+\frac{i_c(\beta+\beta_i)+\beta i_b}{\beta_i(\beta i_b - i_c)}(1+c)\right] \quad (21)$$

where the correction term

$$c = K\frac{i_c(\beta-\beta_i)+\beta i_b(2\beta_i-1)}{\beta+\beta_i}$$

has been assumed small, and β has been assumed large compared with unity. In this expression K is the current-amplification falloff factor as given by equation (9). It should be noted that ohmic resistances of the collector and emitter are usually negligible and hence the saturation voltage given by equation (21) should be quite accurate. However, in calculating the individual junction voltages, the i_bR_b voltage may be dominant and should always be taken into account.

Low-level switching operation

One of the most important uses of transistors under the conditions of the present approximation is that of a low-level signal switch. For this purpose

* The saturation voltage of alloy transistors has been analyzed in great mathematical detail by HUANG[16]. Unfortunately, the two most important features of saturation operation were treated in a self-inconsistent manner. In the determination of the internal lateral base current, the effects of non-unity emitter efficiency were neglected, yet were included later in the calculation of the total base current. Clearly if this contribution were at all important in the total base current, it would be the dominant source of lateral base current. The second inconsistency occurs in the solution for the radial distribution of carrier density assuming the collector and base radii to be equal (a case which is, in general, uninteresting). The results obtained under these assumptions are then extended to the case where the radii are not equal. For any case where the crowding is at all severe this extension is grossly in error. In the light of these errors the value of HUANG's analysis is seriously questionable.

a transistor is universally used in the inverted connection because of the lower saturation voltage obtainable. In this application the emitter current is very small and the transistor is used merely to transfer voltage information from collector to emitter. If the transistor is to operate with any reasonable success in such a connection, ρ_1 and ρ_2 must be very nearly equal, and we may greatly simplify the analysis. From the linear lumped model, the saturation voltage under these conditions is predicted to be

$$v_{\text{sat}} \simeq \frac{kT}{q\beta}$$

However, such operation is quite sensitive to the crowding or self-bias cutoff effect. If we assume $\rho_1 \simeq \rho_2 \simeq \rho$ and $i_e \ll i_b$ the base current may be written

$$i_b \simeq \rho G_{d_0}\left(\frac{1}{\beta}+\frac{1}{\beta_i}\right)$$

The saturation voltage may be evaluated from equation (21), interchanging the inverse and normal parameters, neglecting the i_e terms and assuming $\beta \gg 1$.

$$v_{\text{sat}} \simeq \frac{kT}{q}\ln\left(1+\frac{1+c}{\beta}\right) \tag{22}$$

where

$$c \simeq \frac{2K\beta\beta_i i_b}{\beta+\beta_i}$$

Since it is desired to obtain the lowest-possible saturation voltage, operation at low base currents is seen to be desirable. However, a lower limit is imposed by the junction depletion-layer recombination which causes G_1 to increase at low current levels and therefore results in an increased saturation voltage.

Operation in the region where depletion-layer recombination is important is not recommended for another very important reason. The effective G_1 due to this effect, and hence the effective β, is very dependent upon temperature. Hence the saturation-voltage stability with temperature at small drive currents may be much worse than at higher drive currents.

In any transistor which is at all suitable for this type of operation and which is driven with reasonably small base current, the second term in equation (22) will be small compared to unity. Thus the logarithm may be expanded in terms of its argument, and equation (22) becomes

$$v_{\text{sat}} \simeq \frac{kT}{q\beta}\left(1+\frac{2K\beta\beta_i i_b}{\beta+\beta_i}\right) \tag{23}$$

where again the correction term has been assumed small compared with unity.

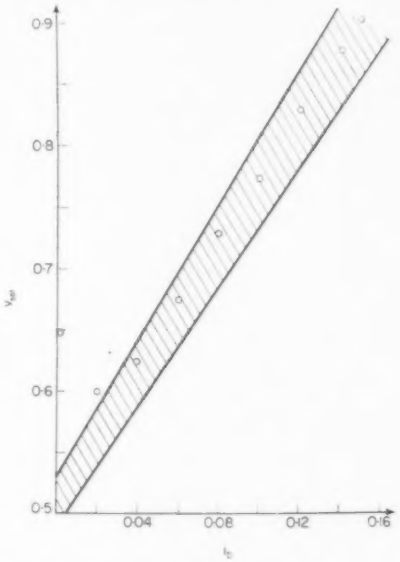


FIG. 7. Experimental verification of equation (23).

Experimental values of v_{sat} for an alloy transistor operating at low level are shown in Fig. 7. The theoretical values fall somewhere within the limits shown, depending upon the exact values of β and β_i used in the calculation. The analysis is especially critical to the precise value of β_i used. Perhaps the most satisfactory method of obtaining β_i and K is to fit the i_e versus i_b data to Fig. 3 and use the normalization constants to determine these parameters. The true low-level value of β_i is often masked by depletion-layer recombination at low operating currents. This effect is quite noticeable in the experimental data of Fig. 7, where the saturation voltage increases slightly at low base currents.

The general requirements on a transistor to be used in low-level switching service are that the saturation voltage should be as independent as possible of time, temperature and variations in

drive current. Thus, for efficient operation in this type of application, the transistor should have a very high β and low value of K. Also the depletion-layer recombination should not become important until very low levels are reached.

An obvious extension of the transistor signal switch is to operate two such switches back-to-back so that their saturation voltages tend to cancel. This procedure has often been used with some success. The problems involved are primarily those of matching β between units, although the variation of v_{sat} with drive current is often objectionable and a matching of the drive correction factor may also be attempted, with the obvious increase in labor involved. Again it should be noted that the depletion-layer recombination rate is extremely variable between units, and hence operation in the region where this condition is important is even more to be avoided.

The potential applications of transistor low-level signal switches are very numerous and exciting. In nearly all such applications the requirements are sufficiently critical that matched back-to-back pairs are desirable. It would seem highly desirable that optimally designed matched units in the same package, mounted on the same mounting base, should be made available to the circuit designer in the same way that avalanche reference elements have been in the past. Perhaps one reason for the overlong delay is the general lack of understanding of the principles underlying the operation of devices used in this type of service.

Very high level

If we attempt to solve for the saturated density distribution under very high-level conditions by returning to the initial integral equations, we are faced with an extremely difficult problem. The introduction of the second variable ρ_2 introduces tremendous complications into the mathematical manipulation and renders a closed-form solution impossible. However, before giving up all hope and resorting to numerical methods, let us consider the real problem with which we are faced. We have derived expressions for the saturation voltage under the assumption that the lumped-model elements were linear, and have seen how correction terms were added even at low levels due to the self-bias effect. However, as we have noted, in transistors designed to eliminate the density crowding (i.e. those with thin emitter stripes) the injected density may well become comparable with n_n at current levels where G_d has not decreased appreciably below its low-level value. In switching service such a unit will nearly always be used in the very high-level region and we are interested in determining the first-order correction to the linear theory under these conditions. As an approximation we will use the variation of G_d with ρ as given in equation (19). Since this variation of G_d with ρ is of the same form as in the low-level case, we will assume the previous analysis applies, and may use the results with the following modifications.

(1) In units where ρ becomes comparable with n_n at reasonably low currents, the value of β as measured or given by the manufacturer is normally twice the low-level value due to the drift enhancement of G_d mentioned earlier. Since at higher levels this effect is still present, we do not need to add any additional factors to the analysis, but need only be sure that the value of β has been measured at a high enough current that the drift enhancement has become effective.

(2) An additional term must be added to the saturation voltage due to the electric field in the base region caused by the gradient of majority-carrier density. Under high-level conditions the total potential drop across the base region is given by the condition that the drift and diffusion tendencies of majority-carrier current are equal.

$$\mu n E = D \frac{dn}{dy}$$

This electric field, when integrated across the base region results in a potential

$$V = \frac{kT}{q} \ln \frac{\rho_1}{\rho_2} \tag{24}$$

Since this voltage is identical to the difference in forward-bias voltages of the junctions and of the same polarity, the effect of high injection level is to double the observed saturation voltage over that given in the low-level case.

(3) It should again be emphasized that the present analysis has not included the effects of electron–hole scattering. The decrease in G_d due to this effect will become appreciable at extremely

high levels, and additional saturation voltage is to be expected.

The results of this paper are quite significant to the circuit engineer in that they permit the first-order characterization of transistor switches by the use of only one additional parameter, i.e. K the current-amplification falloff factor. This modification to the basic theory permits calculations of many switching problems of interest to engineering accuracy. In addition, the author has endeavored to present a clear physical insight into the fundamental limitations of the device so necessary for complex circuit work.

REFERENCES

1. J. G. Linvill, *Proc. I.R.E.* **46,** 1141 (1958).
2. C. A. Mead, Transistor switching analysis, *Semiconductor Products*. To be published.
3. J. L. Moll, *Proc. I.R.E.* **46,** 1076 (1958).
4. C. T. Sah, R. N. Noyce and W. Shockley, *Proc. I.R.E.* **45,** 1229 (1957).
5. W. Shockley and W. T. Read, Jr., *Phys. Rev.* **87,** 835 (1952).
6. W. M. Webster, *Proc. I.R.E.* **42,** 914 (1954).
7. R. N. Hall, *Phys. Rev.* **87,** 387 (1952).
8. R. N. Hall, *Proc. I.R.E.* **40,** 1512 (1952).
9. R. D. Middlebrook, *An Introduction to Junction Transistor Theory*. John Wiley, New York (1957).
10. N. H. Fletcher, *Proc. I.R.E.* **43,** 551, 1669 (1955).
11. D. H. Looney, *Base Resistance Considerations in Transistor Design*, Chap. 8, p. 76. Bell Telephone Labs., 14th Interim Technical Report on Transistor Services (1955).
12. K. E. Mortenson, *I.R.E. Trans. Electron Devices* ED-**6,** 174 (1959).
13. N. H. Fletcher, *Proc. I.R.E.* **45,** 862 (1957).
14. A. R. Moore and J. I. Pankove, *Proc. I.R.E.* **42,** 907 (1954); K. F. Stripp and A. R. Moore, *Ibid.* **43,** 856 (1955).
15. M. Clark, *Proc. I.R.E.* **46,** 1185 (1958).
16. C. Huang, *I.R.E. Trans. Electron Devices* ED-**6,** 141 (1959).
17. H. G. Rudenberg, *Proc. I.R.E.* **46,** 1304 (1958).

See also:

E. S. Rittner, *Phys. Rev.* **94,** 1161 (1954).
W. Shockley, *Bell Syst. Tech. J.* **28,** 435 (1949).
P. G. Griffith, *Lumped Models of Drift Transistors for Large Signals*. Stanford Electronics Laboratories Technical Report No. 1501–1 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 225–233. Printed in Great Britain

ZUR MESSUNG VON MAGNETISCHEN FELDERN MIT HALLGENERATOREN

H. WEISS

Aus dem Forschungslaboratorium der Siemens-Schuckertwerke AG, Erlangen, Germany

(*Received* 4 *May* 1960)

Zusammenfassung—Untersuchungen von Hallgeneratoren aus InAs haben gezeigt, dass die Empfindlichkeit bei der Messung niederfrequenter Magnetfelder im wesentlichen durch das thermische Rauschen des Innenwiderstandes des Halbleiters bestimmt wird, d.h. für das kleinste nachweisbare Feld ist die Halleistung gleich der Rauschleistung $kTdf$. Bei Gleichfeldern ist die Empfindlichkeit um zwei bis drei Grössenordnungen geringer. Dauerversuche über 8 Monate ergaben für Hallgeneratoren aus InAs eine Reproduzierbarkeit der Hallspannung von $\pm 2 \times 10^{-5}$.

Für InSb und InAs mit $N_D = 6 \times 10^{16}/\text{cm}^3$ liess sich unterhalb 10 000 Gauss keine Abhängigkeit des Hallkoeffizienten von der magnetischen Induktion nachweisen.

Abstract—Experiments with InAs Hall generators have shown that, for measuring low-frequency magnetic fields, the sensitivity is determined especially by the thermal noise of the internal resistance of the semiconductor, that is for the smallest detectable field the Hall power is equal to the noise power $kTdf$. For constant magnetic fields the sensitivity is smaller by two to three orders of magnitude. Experiments over a period of 8 months showed a reproducibility of $\pm 2 \times 10^{-5}$ for the Hall voltage of a generator made of InAs.

For InSb and InAs with $N_D = 6 \times 10^{16}/\text{cm}^3$ below 10,000 gauss no dependence of the Hall coefficient on the magnetic induction was found.

Résumé—Des expériences avec des générateurs de Hall de InAs montrent que, pour la mesure des champs magnétiques de basse fréquence, la sensibilité est déterminée spécifiquement par le bruit thermique de la résistance interne des semiconducteurs, c'est-à-dire pour les champs les plus faibles détectables, la puissance de Hall est égale à la puissance de bruit $kTdf$. Pour des champs magnétiques constants la sensibilité est inférieure par 2 ou 3 ordres de grandeur. Des expériences sur une période de 8 mois montrent une reproductibilité de $\pm 2 \times 10^{-5}$ pour la tension de Hall d'un générateur fait de InAs.

Pour InSb et InAs avec $N_D = 6 \times 10^{16}/\text{cm}^3$ au-dessous de 10 000 gauss on ne trouve aucune relation entre le coefficient de Hall et l'induction magnétique.

Die Messung von Magnetfeldern mit Hilfe des Halleffektes hat durch die Auffindung der halbleitenden Verbindungen InSb und InAs mit hoher Elektronenbeweglichkeit grosse Bedeutung erlangt. Die technischen Bauelemente, die diesen Effekt ausnützen, werden Hallgeneratoren genannt.

Sie werden bereits in verschiedenen Gebieten der Elektrotechnik zur Messung, Regelung oder Steuerung verwendet.[1] Die Eigenschaften, die von ihnen gefordert werden, sind nicht immer dieselben und hängen von der besonderen Anwendung ab. In diesem Aufsatz werden Untersuchungen bezüglich der Empfindlichkeit bei der Messung magnetischer Felder mit Hallgeneratoren mitgeteilt.

Es ist bekannt, dass der Wirkungsgrad eines Hallgenerators bei Anpassung allein durch die Elektronenbeweglichkeit des Halbleiters bestimmt wird.[2] Neben einer grossen Elektronenbeweglichkeit als wichtigstem Parameter muss ein idealer Hallgenerator noch weitere Eigenschaften besitzen, die teilweise auch mit einer hohen Elektronenbeweglichkeit verknüpft sind, wie später gezeigt wird. Die geforderten Eigenschaften sind folgende:

(1) Die Nullspannung, d.h. die Spannung

zwischen den Hallsonden ohne Magnetfeld, ist unabhängig von Temperatur und Zeit.

(2) Die Hallspannung im Magnetfeld ist gross und hängt nur schwach von der Temperatur ab.

(3) Die Hallspannung ist eine genau lineare Funktion der magnetischen Induktion.

Von diesen drei Forderungen ist die erste die wichtigste, da die Schwankungen der Nullspannung die Nachweisgrenze für Magnetfelder bestimmen. Die Absolutgrösse der Hallspannung spielt demgegenüber nur eine untergeordnete Rolle. Nichtlinearität und Temperaturabhängigkeit des Hallkoeffizienten stören bei vielen technischen Anwendungen nicht, ausserdem können sie oft durch geeignete Schaltung, z.B. Gegenkopplung, korrigiert werden. Nullspannungsschwankungen des Hallgenerators lassen sich jedoch nicht durch äussere Schaltungsmittel eliminieren.

KONSTANZ DER NULLSPANNUNG

Die Hallsonden geben eine lediglich durch die Schwankungen der Nullspannung bedingte Wechselleistung L_{Sch} ab, wenn der zeitliche Mittelwert der Nullspannung verschwindet. (Entweder ist das von vornherein der Fall, oder der Mittelwert wird kompensiert.) Man kann die kleinste, mit einem Hallgenerator noch nachweisbare magnetische Induktion B_{min} so definieren, dass im Felde B_{min} die Hall-Leistung L_H gleich der Schwankungsleistung L_{Sch} ist.

Nun besteht für die Hall-Leistung bei Anpassung in kleinen Magnetfeldern folgende Beziehung:

$$L_H = G\, L_{St}\, (\mu B)^2 \qquad (1)$$

Hierbei ist G ein Geometriefaktor der ungefähren Grösse 1/16, wenn der Halbleiter doppelt so lang wie breit ist, μ die Elektronenbeweglichkeit und L_{St} die durch den Steuerstrom i_{St} im Halbleiter erzeugte Leistung. Man erhält dann für die kleinste noch nachweisbare Indution B_{min} folgende Beziehung:

$$B_{min} = \sqrt{\left(\frac{L_{Sch}}{G \cdot L_{St}}\right)} \frac{1}{\mu} \qquad (2)$$

Hier steht das wichtige Ergebnis, dass bei einer vorgegebenen Grösse der Schwankungsleistung L_{Sch} die Messempfindlichkeit mit steigender Elektronenbeweglichkeit zunimmt.

Die Spannungsschwankungen zwischen den Hallsonden werden durch ein regelloses Hin- und Herwandern der Äquipotentiallinien ohne Magnetfeld verursacht. Gründe für dieses Wandern sind:

(A) Thermisches Rauschen des hallseitigen Innenwiderstandes: $L_{Sch} = kT \,.\, df$ (df = Bandbreite).
Eingesetzt in Gleichung (2) ergibt das:

$$B_{min} = \sqrt{\left(\frac{kTdf}{G\, L_{St}}\right)} \frac{1}{\mu} \qquad (3)$$

Die Elektronenbeweglichkeit ist hier der einzige Parameter des Halbleiters, der B_{min} bestimmt.

(B) Zeitlich nicht konstante Thermospannungen zwischen den Hallsonden [Ursache wie unter (C)]. Sie stören nicht, wenn der Steuerstrom ein Wechselstrom ist.

(C) Temperaturabhängigkeit des spezifischen Widerstandes des Halbleiters. Um die durch den Steuerstrom im Halbleiter erzeugte Wärmeenergie nach aussen abzuführen, muss ein Temperaturgefälle herrschen. Die Temperaturverteilung im Halbleiter wird jedoch auch bei Anordnung des Hallgenerators in einem Thermostaten infolge von Regelschwankungen des Thermostaten oder Änderungen der Umgebungstemperatur nicht konstant bleiben. Das führt infolge der Temperaturabhängigkeit des spezifischen Widerstandes zu hin- und herwandernden Äquipotentiallinien.

Im Gegensatz zum weissen thermischen Rauschen verursachen die unter (B) und (C) genannten Ursachen nur langsame Schwankungen der Nullspannung.

Zur Untersuchung der Schwankungen der Nullspannungen wurden kommerzielle Hallgeneratoren aus InAs, Type FA 24, verwendet. Zur Messung wurde die in Abb. 1 abgebildete Anordnung gewählt. Der Innenwiderstand des Hallgenerators wurde durch einen Transformator auf den Eingangswiderstand von 2000 Ω des folgenden Transistorverstärkers angepasst. Hinter dem Verstärker lag ein Ringmodulator mit einer Steuerspannung variabler Frequenz bis 5000 Hz.

Die Bandbreite des Systems wurde durch den Frequenzgang des folgenden Schreibers bestimmt. Die Empfindlichkeit der Anlage betrug 10^{-9} V/mm auf dem Registrierpapier, die Bandbreite etwa 2 Hz. Die Schwankungen des Verstärkers betrugen

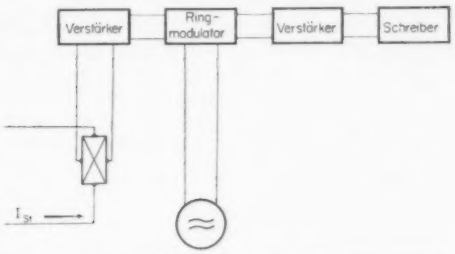


ABB. 1. Blockschaltbild der Verstärkeranordnung zur Messung der Nullspannungsschwankungen.

mit kurzgeschlossenem Eingang 10^{-9} V. Wurde der Hallgenerator angeschlossen, so liess sich keine Änderung der Schwankungen am Ausgang des Verstärkers bei 400, 800, 3000 und 5000 Hz feststellen. Daran änderte sich nichts, wenn der Hallgenerator von einem Steuerstrom (Gleichstrom) von 300 mA, entsprechend einer Steuerleistung von 100 mW, durchflossen war.

Nun entsprechen 10^{-9} V bei einem Steuerstrom von 300 mA einer magnetischen Induktion von 4×10^{-5} G. Setzt man zum Vergleich die experimentellen Daten sowie $\mu = 20\,000$ Vsec/cm^2 in Gleichung (3) ein, so erhält man für eine dem thermischen Rauschen äquivalente Induktion: $B_{min} = 6 \times 10^{-6}$ G. Daraus folgt, dass die Spannungsschwankungen des Hallgenerators zwischen 400 und 5000 Hertz sicher nicht wesentlich grösser als das thermische Rauschen sind, und dass daher dieses die Empfindlichkeit bei der Messung niederfrequenter Magnetfelder bestimmt.

Die experimentellen Untersuchungen der Schwankungen im Gebeit von einigen Hertz und darunter erfolgten ebenfalls mit der Anordnung nach Abb. 1. Der Steuerstrom war ein Wechselstrom von 500 Hz und wurde von derselben Spannungsquelle geliefert wie die Steuerspannung des Modulators. Der Hallgenerator befand sich zwischen zwei Kupferplatten, die durch aufgelötete Kupferrohre, durchflossen vom Wasser eines Thermostaten, auf konstanter Temperatur gehalten wurden. Betrug die Messdauer einige Minuten, so entsprachen die Schwankungen einer Feldstärke von einigen 10^{-4} G, unabhängig von der Grösse des Steuerstromes; Registrierung der Hallspannung während eines Tages vom Morgen bis zum Abend ergab Schwankungen entsprechend $\pm 2 \times 10^{-3}$ G.

Wurde der Hallgenerator mit Gleichstrom betrieben und die Hallspannung mit einem lichtelektrischen Verstärker auf den Schreiber gegeben, so erhielt man die gleichen Ergebnisse.

Diese Versuche zeigen, dass bei Messungen über längere Zeiten die Empfindlichkeit des Hallgenerators durch die Temperaturschwankungen im Innern des Halbleiters um zwei Grössenordnungen kleiner ist als allein durch thermisches Rauschen bei Verstärkung mit kleiner Bandbreite. Änderungen der Thermospannung zwischen den Hallsonden sind demgegenüber gering. Die Empfindlichkeitsgrenze lässt sich dann noch um zwei bis drei Grössenordnungen herabdrücken, wenn man den Halbleiter zwischen zwei das Magnetfeld konzentrierende Stäbe aus Weicheisen setzt.[3] Damit wird eine Nullpunktskonstanz über mehrere Stunden erreicht, die einem Feld von 10^{-5} G entspricht.

DAUERVERSUCHE ÜBER 8 MONATE

Bei den folgenden Experimenten wurde die Konstanz von Nullspannung und Hallkoeffizient geprüft. Sie wurden an zwei Hallgeneratoren aus InAs vom Typ FA 24 mit einem Hallkoeffizienten $R_H = 100$ cm^3/Asec durchgeführt. Der eine war eine fertige kommerzielle Ausführung, der andere halbfertig, d.h. die Oberseite des auf einen Träger geklebten Halbleiters war nicht vergossen, sondern befand sich frei an Luft. Es sollte damit ein etwaiger Einfluss der Vergussmasse oder der atmosphärischen Umgebung auf den Halbleiter im Laufe der Zeit festgestellt werden.

Die Versuchsanordnung mit Schaltbild ist in Abb. 2 abgebildet. Die beiden Hallgeneratoren wurden vom gleichen Steuerstrom durchflossen (Abb. 2 oben) und lagen in Reihe mit einem Normalwiderstand N von 1 Ω sowie einem Eisenwasserstoffwiderstand. Ein 6-V-Akkumulator lieferte einen Steuerstrom von etwa 100 mA. Ein konstanter Widerstand wurde anstelle eines veränderlichen verwendet, weil ein solcher infolge des schlechten Kontaktes seines Schleifers unkontrollierbare Stromschwankungen verursacht.

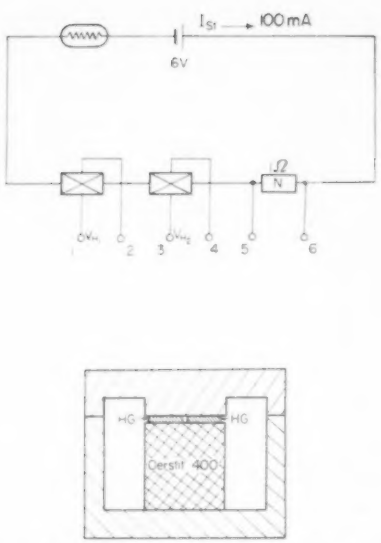


ABB. 2. Oben: Schaltbild für Dauerversuche; Unten: Anordnung der Hallgeneratoren *HG* im Topfmagneten.

Die Kontaktpaare 1,2 und 3,4, die mit den Hallsonden der beiden Hallgeneratoren verbunden waren, sowie die Anschlüsse 5,6 am Normalwiderstand konnten wahlweise mit dem Eingang eines Diesselhorst-Kompensators verbunden werden. Bei einer Hallspannung von etwa 50 mV betrug die Ablesegenauigkeit am Nullinstrument $\pm 0{,}5\,\mu$V. Die Spannungen an den Hallgeneratoren und am Normalwiderstand wurden mehrere Male abwechselnd gemessen, um sich zu vergewissern, dass während des Umschaltens von den Klemmen 1,2 bzw. 3,4 auf 5,6 der Steuerstrom sich nicht änderte. Die auf einen Steuerstrom von 100 mA bezogenen Hallspannungen waren dann mit einer Genauigkeit von $\pm 2 \times 10^{-5}$ bestimmt.

Die beiden Hallgeneratoren *HG* waren nebeneinander in demselben Luftspalt eines Topfmagneten angeordnet (Abb. 2 unten). Der magnetische Kreis bestand aus einem Dauermagneten aus Oerstit 400 und einem Topf aus Weicheisen. Die magnetische Induktion im Luftspalt betrug etwa 6000 G. Die Messgenauigkeit entsprach dann einer Feldänderung von $\pm 0{,}1$ G. Der Dauermagnet befand sich in einem abgeschlossenen Behälter mit aussen angelöteten Kupferrohren. Diese wurden von Leitungswasser durchflossen, dessen Temperatur lediglich langsame Schwankungen von etwa 5° C im Laufe eines Jahres aufwies. Sie wurde mit einem Quecksilberthermometer auf $\pm 0{,}1°$ C bestimmt.

Die Wahl der Anordnung der beiden Proben hat folgenden Grund: Wenn sich ein Hallgenerator im Felde eines Dauermagneten befindet, so gibt es zwei Ursachen für die Änderung der Hallspannung:

(1) Änderung der Induktion im Luftspalt durch Schwankungen der geometrischen Dimensionen des Spaltes und des magnetischen Momentes des Dauermagneten, vor allem infolge von Temperaturänderungen, sowie durch äussere Streufelder. Diesen Einflüssen unterliegen die beiden Hallgeneratoren in gleicher Weise.

(2) Änderungen der individuellen Eigenschaften der einzelnen Hallgeneratoren. Diese sind unabhängig voneinander.

Die Aufgabe bestand nun darin, die unter 2. genannten Schwankungen festzustellen.

Der Topfmagnet war 8 Monate lang verschlossen. An allen in diesem Zeitraum liegenden Werktagen wurden die Spannungen an den beiden Hallgeneratoren gemessen. Die Temperatur war im Sommer mit 14°C am höchsten und sank im Winter bis auf 9° C ab.

AUSWERTUNG DER MESSERGEBNISSE

Bei der ersten Auswertung wurden die beiden Hallgeneratoren völlig unabhängig voneinander behandelt, d.h. so, als ob es nur individuelle Fehler gäbe. Dazu wurden jeweils die in den 8 Monaten erhaltenen 192 Messwerte der Hallspannung verwendet. Diese sind für einen Hallgenerator in Abb. 3 in Abhängigkeit von der Temperatur aufgetragen. Mit Hilfe der Methode der kleinsten Quadrate wurde dann eine wahrscheinlichste Gerade für die lineare Beziehung zwischen Hallspannung und Temperatur berechnet. Sie ist in Abb. 3 die mittlere der drei Geraden. Ihre Neigung ist grösser als für den Hallgenerator allein, da sie sowohl den Temperaturgang des Hallgenerators als auch den der magnetischen Induktion im Luftspalt umfasst. Aus den Abweichungen der Messpunkte von der wahrscheinlichsten Geraden lässt sich der mittlere Fehler der Einzelmessung bestimmen. Die Auswertung der

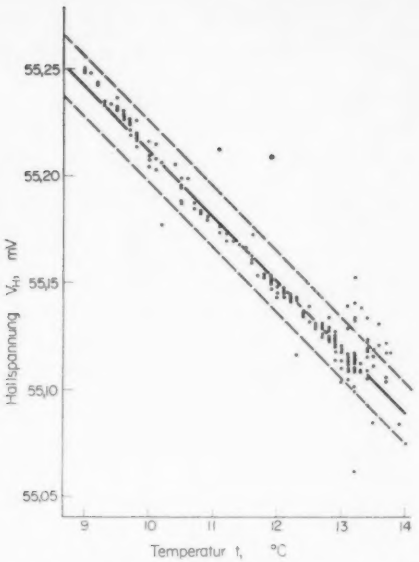


ABB. 3. Ergebnis der Dauerversuche über 8 Monate mit einem Hallgenerator.

in den 8 Monaten erhaltenen 2×192 Messpunkte ergab folgende relative mittlere Fehler:

$$\text{Probe 1: } 3{,}2 \times 10^{-4} \quad \text{Probe 2: } 3{,}5 \times 10^{-4}$$

Einem relativen Messfehler von $\pm 3{,}3 \times 10^{-4}$ entsprechen die gestrichelten Geraden in Abb. 3.

Bei der zuletzt angegebenen Auswertung war angenommen worden, dass die Fehler in den Messungen der beiden Hallgeneratoren völlig unabhängig voneinander sind. Das ist jedoch keineswegs der Fall. Das zeigt Abb. 4, in der die Hallspannungen der beiden Hallgeneratoren in Abhängigkeit von der Temperatur für den Monat März dargestellt sind: Die Schwankungen der Messwerte sind ungefähr gleich für beide Messreihen. Das lässt darauf schliessen. dass die Abweichungen der Messpunkte von einer Geraden für beide Hallgeneratoren im wesentlichen gemeinsame Ursachen haben, die individuellen Schwankungen jedoch dagegen klein sind. Für ein gemeinsames Streuen der beiden Messreihen gibt es neben den unter (1). oben genannten Gründen noch einen anderen messtechnischen:

Die den beiden Hallgeneratoren gemeinsame Temperatur ist falsch bestimmt worden. Das kann daher kommen, dass das Thermometer ausserhalb des Topfmagneten angebracht ist und die Ablesung nur auf $\pm$ 0,1° genau ist.

Um die individuellen Schwankungen der Hallspannung zu ermitteln, muss man die gemeinsamen regelmässigen Fehler absondern. Dazu dient das

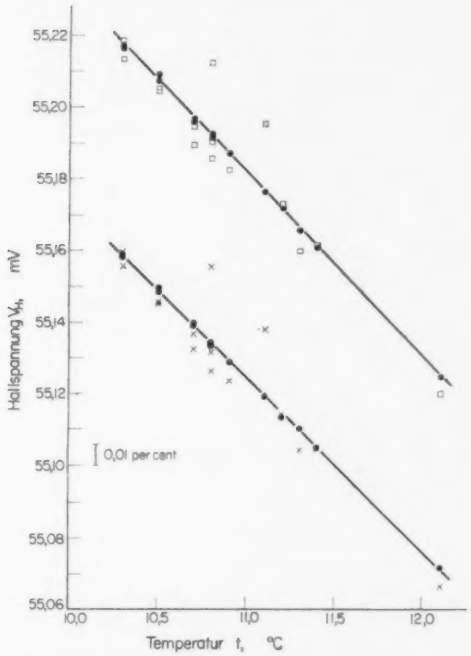


ABB. 4. Ergebnisse der Dauerversuche über einen Monat mit 2 Hallgeneratoren: x, □ Messwerte, ● die um den regelmässigen Fehler korrigierten Messwerte.

im folgenden beschriebene Verfahren, das allgemein bei der gleichzeitigen Messung von n Hallspannungen zur Trennung der regelmässigen und der zufälligen Fehler dienen kann.

$U_{\gamma\mu}$ sei die am Tage μ bei der Temperatur t_μ gemessene Hallspannung des γ-ten Hallgenerators.

Für die mit einem Stern (x) bezeichneten wahren Werte von U und t soll eine lineare Beziehung bestehen:

$$U^x_{\gamma\mu} = a_\gamma + b_\gamma t^x_\mu,\ \gamma = 1, \ldots n,\ \mu = 1, \ldots m \quad (4)$$

Die bei der μ-ten Messung erhaltene Hallspannung $U_{\gamma\mu}$ setzt sich aus dem wahren Wert $U^x_{\gamma\mu}$, aus dem zufälligen Fehler $z_{\gamma\mu}$ und aus dem regelmässigen

Fehler $r_{\gamma\mu}$ zusammen:

$$U_{\gamma\mu} = U^x_{\gamma\mu} + z_{\gamma\mu} + r_{\gamma\mu} \tag{5}$$

Die Messung der Temperatur besitzt, da sie für die n Halbleiter gleichzeitig erfolgt, nur einen von γ unabhängigen regelmässigen Fehler s_μ.

$$t_\mu = t^x_\mu + s_\mu \tag{6}$$

Da alle $U_{\gamma\mu}$ etwa die gleiche Grösse haben, lassen sich die regelmässigen Fehler $r_{\gamma\mu}$ proportional zum wahren Wert $U^x_{\gamma\mu}$ ansetzen, wobei der Faktor unabhängig von γ ist; das gilt sicher, wenn Feld und Temperatur schwanken.

$$r_{\gamma\mu} = \Delta_\mu U^x_{\gamma\mu} \tag{7}$$

In dieses Δ_μ kann man auch den Fehler s_μ in der Bestimmung von t_μ stecken, so dass man bei der weiteren Rechnung $t_\mu = t^x_\mu$ annehmen kann. Aus Gleichungen (5) und (7) ergibt sich dann:

$$U_{\gamma\mu} = U^x_{\gamma\mu}(1+\Delta_\mu) + z_{\gamma\mu} \tag{8}$$

Summation über γ liefert:

$$\sum_{\gamma=1}^{n} U_{\gamma\mu} = (1+\Delta_\mu) \sum_{\gamma=1}^{n} U^x_{\gamma\mu} + \sum_{\gamma=1}^{n} z_{\gamma\mu} \tag{9}$$

Die Messwerte $U_{\gamma\mu}$ sind nahezu gleich und die zufälligen Fehler unabhängig voneinander; es ist also sinnvoll die Summe $\sum_{\gamma=1}^{n} z_{\gamma\mu}$ gleich Null zu setzen. $\sum_{\gamma=1}^{n} U_{\gamma\mu}$ ist dann die mittlere Summe der Einzelmessungen für ein bestimmtes μ.

Da nun $U^x_{\gamma\mu}$ für jedes γ eine lineare Funktion von t_μ ist, muss auch die Summe $\sum_{\gamma=1}^{n} U^x_{\gamma\mu}$ eine lineare Funktion von t_μ sein. Trägt man die nach Gleichung (9) ermittelte Summe der Einzelmessungen in Abhängigkeit von t auf, so wird man keine genau lineare Funktion erhalten, sondern die Punkte liegen um eine die wahre Summe der Hallspannungen darstellende Gerade gestreut. Der Abstand der einzelnen Punkte von dieser Geraden ist durch Δ_μ gegeben. Man kann dann für die mittlere Summe mit Hilfe der Methode der kleinsten Quadrate eine wahrscheinlichste Gerade berechnen. Die Abweichungen der mittleren Summen von dieser wahrscheinlichsten Geraden liefern den wahrscheinlichsten regelmässigen Fehler Δ'_μ der Einzelmessung. Statt Gleichung (8) hat man dann:

$$U_{\gamma\mu} = U^x_{\gamma\mu}(1+\Delta'_\mu) + z'_{\gamma\mu}.\ z'_{\gamma\mu} \approx z_{\gamma\mu} \tag{10}$$

Da $\Delta_\mu \ll 1$ ist, kann man auch schreiben:

$$U_{\gamma\mu}(1-\Delta'_\mu) = U'_{\gamma\mu} = U^x_{\gamma\mu} + z'_{\gamma\mu} \tag{11}$$

$U'_{\gamma\mu}$ ist der um den wahrscheinlichsten regelmässigen Fehler korrigierte Messwert. Man ermittelt jetzt für die jeweils zu einem γ gehörenden korrigierten Messwerte mit der Methode der kleinsten Quadrate die wahrscheinlichsten Werte für a_γ und b_γ. Die Abweichungen der $U'_{\gamma\mu}$ von der so errechneten wahrscheinlichsten Geraden geben die zufälligen Fehler der einzelnen Messung an. Daraus lässt sich nach den bekannten Formeln der mittlere zufällige Fehler der Einzelmessung berechnen.

Nun ist der Unterschied zwischen Δ_μ und Δ'_μ um so geringer, je grösser die Zahl n der gleichzeitigen Beobachtungen ist. Die Rechnung nach der angegebenen Methode liefert jedoch auch schon für $n = 2$ eine Abschätzung des mittleren zufälligen Fehlers, da der regelmässige Fehler erheblich überwiegt.

In Abb. 4 sind nun neben den gemessenen Werten auch die um den wahrscheinlichsten regelmässigen Fehler korrigierten Messwerte (dargestellt durch volle Kreise) eingetragen. Man sieht deutlich, dass sie kaum mehr von der durch die Ausgleichsrechnung ermittelten wahrscheinlichsten Geraden abweichen. Dementsprechend ist der mittlere zufällige Fehler der Einzelmessung sehr gering. Er ist in der Tabelle 1 für vier Monate angegeben.

Tabelle 1. *Mittlerer Fehler der Einzelmessung*

Monat	
1	$1{,}4 \times 10^{-5}$
4	$1{,}4 \times 10^{-5}$
6	$1{,}1 \times 10^{-5}$
8	$1{,}7 \times 10^{-5}$

Bie den vier ausgewerteten Monaten ist der mittlere Fehler nicht grösser als die Messgenauigkeit von $\pm 2 \times 10^{-5}$.

Damit wurde gezeigt, dass die Nullspannung

des Hallgenerators im Laufe von 8 Monaten weniger schwankt als der Hallspannung im Felde von 0,1 G entspricht und zugleich der Hallkoeffizient mit der Genauigkeit $\pm 2 \times 10^{-5}$ erhalten bleibt.

GRÖSSE DER HALLSPANNUNG UND IHRE TEMPERATURABHÄNGIGKEIT

Es wurde eingangs erwähnt, dass der Wirkungsgrad allein von der Elektronenbeweglichkeit abhängt. Demnach erscheint Wismut mit $\mu = 5000$ cm²/Vsec für einen Hallgenerator besser geeignet als Germanium. Infolge des geringen spezifischen Widerstandes von Bi sind Hallspannung und Innenwiderstand eines Hallgenerators aus Bi jedoch so klein, dass eine Anpassung an Messinstrumente und Verstärker und damit eine Ausnützung des maximalen Wirkungsgrades nicht mehr möglich ist. Zwischen Hallspannung V_H, Steuerstrom i_{St} und Hallkoeffizient R_H besteht folgende Beziehung:

$$V_H = \frac{R_H \, i_{St}}{d} B \qquad (12)$$

Betrachtet man die Absolutgrösse der Hallspannung, so ist es zweckmässig, in Gleichung (12) den Steuerstrom durch die dem Halbleiter zugeführte Steuerleistung L_{St} und den Innenwiderstand R_i bzw. R_H/μ zu ersetzen, da der Steuerleistung im allgemeinen enge Grenzen zwischen 0,1 und 1 Watt gesetzt sind. Man erhält dann:

$$V_H = \sqrt{\left(\frac{L_{St} \, R_H \, \mu}{2d}\right)} B \qquad (13)$$

$$= \tfrac{1}{2} \sqrt{(R_i \, L_{St})} \, \mu B$$

Bei vorgegebenen Werten von B und L_{St} weist derjenige Hallgenerator die grösste Spannung auf, der das grösste Produkt $R_H \mu / d$ besitzt. Nun kann man die Dicke wegen der mechanischen Festigkeit und wegen der Herstellungsmöglichkeiten nicht beliebig klein machen. Ausserdem nimmt bei sehr kleinen Schichtdicken die Elektronenbeweglichkeit ab, so dass die untere Grenze von d bei einigen 10^{-3} mm liegt. Um eine möglichst hohe Hallspannung zu erhalten, muss man daher ein Material suchen, bei dem mindestens eine der beiden Grössen R_H und μ gross ist. Da nach dem vorigen Abschnitt die Genauigkeit der Messung mit zunehmender Elektronenbeweglichkeit wächst, wird man immer einen Halbleiter mit einem möglichst grossen Wert von μ verwenden. Darüber hinaus kann man den Hallkoeffizienten so gross wählen, als es Bandbreite und Temperaturabhängigkeit erlauben. Je grösser R_H ist, um so kleiner ist bei gleicher Steuerleistung und Beweglichkeit der Steuerstrom.

Ein Halbleiter kann einen um so höheren, wenig von der Temperatur abhängenden Hallkoeffizienten besitzen, je grösser die Breite der verbotenen Zone ist. Abb. 5 zeigt den Zusammenhang zwischen R_H und der Mindestbandbreite $\Delta E_{\min}$, die ein

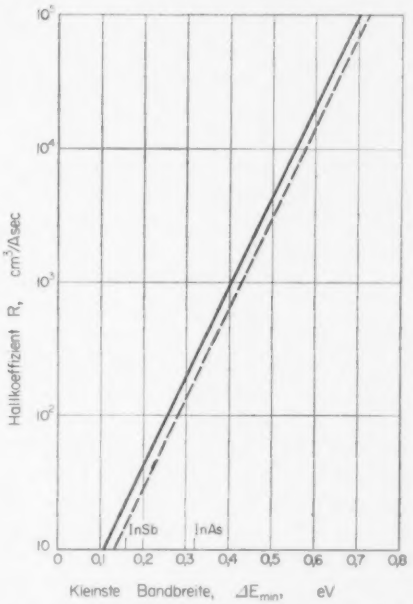


ABB. 5 R_H in Abhängigkeit von $\Delta E_{\min}$ bei 100°C, $(m_n m_p / m_0^2)^{1/2} = 0{,}1$; ——— Abnahme von R_H 10%; – – – – – – – – – Abnahme von R_H 5%.

Material bei 100° C besitzen muss, wenn R_H beim Erwärmen auf 100° C von Zimmertemperatur nicht mehr als 10 Prozent bzw. 5 Prozent abnehmen darf. Die Aktivierungsenergie der Störstellen wurde dabei gleich Null gesetzt; die Kurven gelten auch für den Fall, dass bei nicht verschwindender Aktivierungsenergie alle Störstellen ionisiert sind. Für die Berechnung der Kurven wurde die mittlere relative effektive Masse $(m_n m_p / m_0^2)^{1/2}$ gleich 0,1 gesetzt. Dieser Wert entspricht etwa den bei InSb und InAs gefundenen Zahlen[4] und

ist typisch für einen Halbleiter mit hoher Elektronenbeweglichkeit.

Um eine erheblich höhere Hallspannung, als mit InSb oder InAs möglich, zu erreichen, muss man einen Halbleiter mit einer wesentlich grösseren Bandbreite verwenden, da die Elektronenbeweglichkeit im allgemeinen mit wachsendem ΔE abnimmt. Da es ausserdem immer schwieriger wird, nicht sperrende, nahezu punktförmige Kontakte herzustellen, je grösser Bandbreite und Hallkoeffizient des Halbleiters werden, besitzt InAs mit $\Delta E = 0{,}32$ eV bei 100° C eine sehr günstige Breite der verbotenen Zone. Für die Herstellung eines Hallgenerators hoher Hallspannung sind InP und GaAs dem Si prinzipiell überlegen, da sie neben einer höheren Bandbreite erheblich höhere Elektronenbeweglichkeiten besitzen.

LINEARITÄT DER HALLSPANNUNG

Eine genaue Linearität zwischen Hallspannung und magnetischer Induktion ist keineswegs allgemein, da es bei jedem Halbleiter Dotierungsgebiete gibt, in denen V_H nicht proportional zu B ansteigt. Es ist sogar möglich, dass V_H im Felde von einigen tausend Gauss das Vorzeichen wechselt.[5] Ausserdem spielt die Orientierung der Kristalle eine Rolle,[6]. Es erschien daher notwendig, an Einkristallen aus InSb, InAs und InP geeigneter Dotierung die Abhängigkeit des Hallkoeffizienten von der magnetischen Induktion zu untersuchen. Die verwendeten Präparate waren lange, dünne Stäbchen, so dass eine Verfälschung der Messergebnisse durch die Formabhängigkeit der Hallspannung ausgeschlossen war.[7] Die Hallspannung wurde an jeweils zwei Stellen desselben Stäbchens mit spitzen Sonden gemessen, die mit Federn angedrückt waren. Das Magnetfeld war mit Hallgeneratoren kontrolliert, die mit einer ballistischen Kompensationsmethode überprüft wurden. Die Genauigkeit bei der Bestimmung der magnetischen Induktion betrug mindestens $\pm 0{,}2$ Prozent; Hallspannung und Steuerstrom konnten genauer gemessen werden. Abb. 6 zeigt die Ergebnisse der Untersuchung an je einem Kristall aus InSb und InAs zwischen 300 und 10 000 G. Danach beträgt der mittlere Fehler der Einzelmessung nur 0,15 Prozent. Die Tabelle 2 enthält weitere Ergebnisse bezüglich der Linearität der Hallspannung.

Tabelle 2

Substanz	μ (cm²/Vsec)	R_H (cm³/Asec)	Feldbereich (G)	Fehler ± %
InAs	20 000	100	10^{-3}–400	1
			9000–32 000	0,3
InP	4200	1400	1000–10 000	0,5

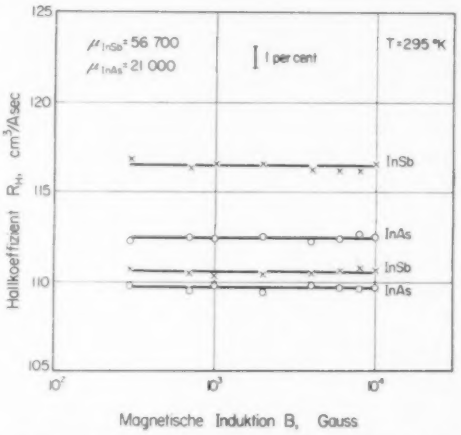


ABB. 6. Hallkoeffizient von InSb und InAs in Abhängigkeit von der magnetischen Induktion.

Die angegebenen Fehler beziehen sich auf den Wert bei der grössten Induktion, sie waren in keinem Falle grösser als der Messgenauigkeit entsprach. Bei den angegebenen Dotierungen und Magnetfeldern liess sich also keine Abhängigkeit des Hallkoeffizienten von der magnetischen Induktion nachweisen. Diese Konstanz von R_H bleibt jedoch bei gelöteten Hallsonden nicht erhalten, sondern die Hallspannung steigt stärker als proportional zum Magnetfeld. Diese Überlinearität kommt andererseits sehr gelegen, da man damit die Zunahme des Innenwiderstandes mit wachsender Induktion B kompensieren kann.[2]

Die Prüfung der Beziehung zwischen Hallspannung und magnetischer Induktion war jedoch nicht auf die in der Tabelle angegebenen Magnetfelder beschränkt, sondern erstreckte sich bis 120 000 G. Bei diesen Induktionen erfolgte die Messung in einer Luftspule, deren Erregerstrom von einer Kondensatorbatterie geliefert wurde.

Bei einem Hallgenerator Type FA 22 war im Bereich von 10–120 kG die Abweichung der Leerlaufspannung von der Linearität kleiner als 2 Prozent.

ZUSAMMENFASSUNG

Die beschriebenen Untersuchungen haben gezeigt, dass Hallgeneratoren aus III–V-Verbindungen infolge ihrer hohen Elektronenbeweglichkeit eine grosse Konstanz der Nullspannung auch über einen Zeitraum von mehreren Monaten und damit eine grosse Messempfindlichkeit besitzen. Dauerversuche mit einem Hallgenerator, dessen Oberfläche der Laboratoriumsluft ausgesetzt war, zeigten, dass das für Hallgeneratoren verwendete InAs infolge des relativ grossen Störstellengehaltes von $6 \times 10^{16}/\text{cm}^3$ sehr unempfindlich gegenüber atmosphärischen Einflüssen ist. Ausserdem besteht bei der angegebenen Dotierung und Verwendung von punktförmigen Hallsonden eine genaue Linearität zwischen Hallspannung und magnetischer Induktion für InAs und InSb.

LITERATUR

1. F. KUHRT, *Siemens-Z.* **28**, 370 (1954).
2. H. WELKER, *Elektro-tech. Z.* **76**, 513 (1955).
3. H. HIERONYMUS und H. WEISS, *Siemens-Z.* **8**, 404 (1957).
4. O. G. FOLBERTH, O. MADELUNG und H. WEISS, *Z. Naturf.* **9a**, 954 (1954).
5. H. WEISS, *Z. Naturf.* **8a**, 463 (1953).
6. W. P. MASON, W. H. HEWITT und R. F. WICK, *J. Appl. Phys.* **24**, 166 (1953).
7. H. J. LIPPMANN und F. KUHRT, *Z. Naturf.* **13a**, 474 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 234–244. Printed in Great Britain

SEMICONDUCTOR TYPE AND LOCAL DOPING DETERMINED THROUGH THE USE OF INFRARED RADIATION

N. J. HARRICK

Philips Laboratories, Irvington-on-Hudson, New York

(*Received* 19 *May* 1960)

Abstract—Infrared techniques for determining semiconductor type and local doping are discussed and examples of measurements as applied to inhomogeneities in germanium are given. The techniques outlined here are generally most sensitive in doping ranges where standard techniques are insensitive. Semiconductor type is determined from the direction of drift of an added carrier packet, thus the minority carrier is determined. Three methods are discussed for determination of semiconductor doping level and its homogeneity. Firstly, a measure of the infrared absorption of the sample gives the equilibrium carrier density from which the doping density can be computed; secondly, a determination of the maximum extraction level can be used to calculate the local sample doping level; thirdly, since any inhomogeneities in a sample act like junctions, local injection–extraction effects occur if a current is passed through the sample, and from observation of these effects the nature of the local doping can be deduced. The first approach for determining sample doping and homogeneity is more useful for heavily doped samples while the latter two are more useful for near-intrinsic samples. The advantages of the techniques outlined here are that sample preparation is particularly simple, sample parameters are not masked by the surface, and information about the nature of the sample deep within the bulk can be obtained without dissecting the sample.

Résumé—On discute des techniques infrarouges de détermination du type de semiconducteurs et du dopage local et on donne des exemples de mesures appliquées à des inhomogénéités dans le Germanium. Les techniques données ici sont, en général, plus sensible dans le domaine où les techniques standard ne sont pas sensibles. La détermination du type de semiconducteurs est faite par une direction d'entrainement d'un paquet de porteurs additionnels, ainsi on détermine les porteurs minoritaires. On discute trois méthodes différentes pour la détermination du niveau de dopage du semiconducteur et de son inhomogénéité.

Tout d'abord, une mesure de l'absorption infrarouge donne la densité des porteurs à l'équilibre à partir de laquelle on peut déterminer la densité de dopage. Ensuite, une détermination du niveau maximum d'extraction peut être utilisé pour calculer le niveau du dopage local. Enfin, puisque toute inhomogénéité agit comme une jonction, on obtient des effets d'injection et d'extraction locales si on passe un courant à travers l'échantillon. De l'observation de cet effet, on déduit la nature du dopage local.

La première méthode pour la détermination des dopages de l'échantillon et de l'inhomogénéité est applicable surtout pour des échantillons très dopés. Alors que les deux dernières sont utilisables surtout pour des échantillons qui sont presque intrinsèques. L'avantage de ces techniques réside dans le fait que la préparation des échantillons est très simple. Les paramètres ne sont pas masqués par la surface et on peut obtenir des informations sur la nature de l'intérieur des échantillons sans que l'on ait à démolir l'échantillon.

Zusammenfassung—Es werden die Infrarot-Techniken zur Bestimmung des Halbleitertyps und der örtlichen Dotierung diskutiert. Beispiele von Messungen werden angeführt, wie sie für Inhomogenitäten in Germanium angewendet werden. Die hier entworfenen Techniken sind im allgemeinen in den Dotierungsbereichen am empfindlichsten, wo die Standard-Techniken unempfindlich sind. Der Halbleitertyp wird nach der Driftrichtung eines zusätzlichen Ladungsträgerpaketes bestimmt, d.h. die Minoritätsträger werden ermittelt. Zur Bestimmung der Störstellenkonzentration des Halbleiters und ihrer Homogenität werden drei Methoden diskutiert: (1) Eine Messung der Infrarotabsorption der Probe ergibt die Gleichgewichtskonzentration der Ladungsträger, nach der die Dotierungskonzentration berechnet werden kann. (2) Eine Bestimmung des

maximalen Extraktionsniveaus kann benutzt werden, um die örtliche Dotierung der Probe zu berechnen. (3) Da Inhomogenitäten in einer Probe ähnlich wie p–n-Übergänge wirken, treten örtliche Injektions–Extraktions-Effekte auf, wenn ein Strom durch die Probe fliesst. Aus der Beobachtung dieser Effekte kann die Natur der örtlichen Dotierung abgeleitet werden. Die erstgenannte Methode zur Bestimmung von Probendotierung und Homogenität ist für hochdotierte Proben besser geeignet, während die beiden letzteren für Proben in der Nähe der Eigenleitung brauchbarer sind. Die Vorteile der hier angeführten Techniken bestehen darin, dass die Präparation der Proben besonders einfach ist, die Parameter der Proben nicht durch Oberflächeneffekte überdeckt werden, und dass Kenntnisse über die inneren Verhältnisse der Probe erlangt werden können, ohne dass die Probe zerschnitten werden muss.

INTRODUCTION

SEMICONDUCTOR type and doping and its homogeneity cannot always be determined easily using standard techniques such as resistivity and thermoelectric-power measurements. This is especially true for near intrinsic semiconductors. One reason for this is that the surface of the semiconductor, which may be of a different type than the bulk, is very high for near-intrinsic semiconductors; e.g. 2·2 Ω-cm per °C for germanium. Thus, other techniques are valuable.

It has already been pointed out that drift-mobility measurements, which yield directly the ambipolar mobility, are especially sensitive to doping density for near intrinsic semiconductors and can be used to determine the doping.[1] Photo-

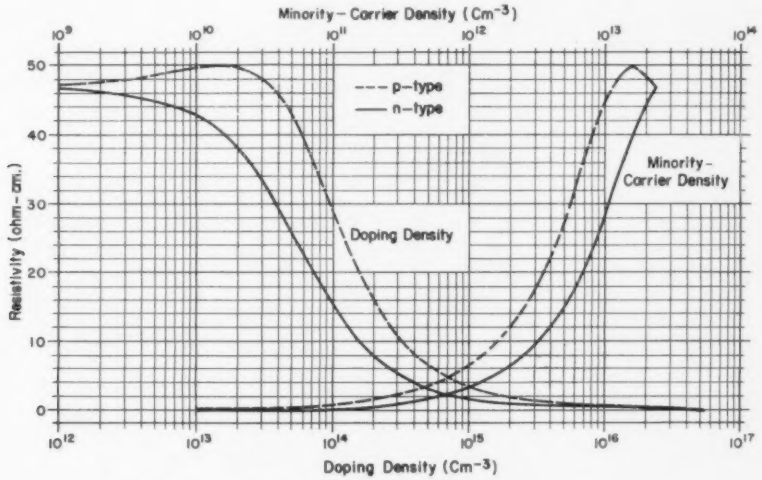


FIG. 1. Carrier densities for germanium vs. resistivity at 300°K.

strongly involved in the usual measurements and may even be the controlling factor. Furthermore, the bulk itself may vary from the outside to the center and this information cannot then be obtained by standard techniques without dissecting the sample. It is evident from Fig. 1 that for near-intrinsic semiconductors resistivity measurements are insensitive to doping density and may even be misleading in the region where the resistivity is double valued as a function of doping. Furthermore, the temperature coefficient of resistivity is voltage measurements have been used successfully to probe out sample inhomogeneities,[2] although it is necessary to dissect the ingots to obtain information about the inhomogeneity deep within the ingot.

Here we describe the ways in which the infrared technique may be used to determine semiconductor type and local doping. So far as measurements of doping density and its homogeneity are concerned, the infrared technique complements resistivity measurements, being especially sensitive in the

near-intrinsic region where resistivity measurements are insensitive or fail completely. In general, the infrared technique has the following advantages over standard techniques:

(a) The sample non-uniformities and type are not masked by the surface.
(b) Inhomogeneities deep within the bulk can be probed out without dissecting the sample.
(c) Generally no elaborate sample preparation is required to make the measurements.

EXPERIMENTAL TECHNIQUE

Fig. 2 shows the experimental set-up for the infrared technique as has been used to investigate carrier distributions in the semiconductor bulk[3] and to deduce properties of a semiconductor surface.[4] The Globar, source of infrared radiation, is focused onto a slit near the sample under investigation. The radiation band width is determined by suitable filters. The radiation transmitted by the sample is detected by a thermopile or other suitable detector. In general, we measure, as a function of position, the detector signals, I_T and ΔI, which are related, respectively, to the degree of transmission of infrared by the sample and changes thereof as the carrier density is changed due to the passage of current through the sample. The ratio of these two quantities gives directly the change in carrier density; i.e.

$$\Delta p = \frac{C}{d}\frac{\Delta I}{I_T}$$

where d is the distance in the sample traversed by the infrared. The various ways in which the calibration constant C can be determined have been enumerated elsewhere.[4] The above simple relationship holds so long as the following conditions are satisfied:

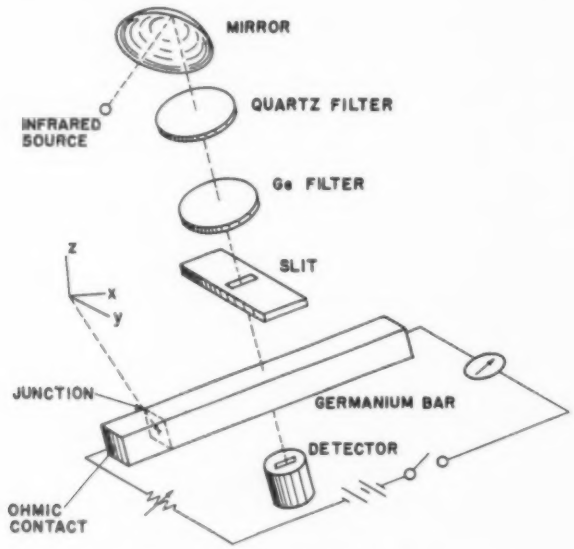


FIG. 2. Schematic diagram of the infrared set-up for measuring carrier distributions in semiconductors.

(a) The injection level is low so that $\Delta I/I_T \ll 1$, otherwise an exponential relationship must be used.
(b) The higher-order transmitted components resulting from multiple reflections are negligible or are eliminated (they can be eliminated if oblique incidence is used).
(c) The response of the detector is linear. This is the case for a thermopile. For a lead sulphide cell a correction must be made if the cell is operated at maximum sensitivity.

(d) The infrared wavelength band must be limited so that there is not a strong wavelength dependence of the absorption coefficient for the free carriers within the band used. In germanium, for example, there is not a strong wavelength dependence of the absorption coefficient in the range 2–5μ. We limit the wavelength band from 2–3·5μ by means of Ge and quartz filters.

(e) The applied electric fields must not be too high so that charge neutrality is maintained and the carrier temperature is not significantly changed above that of the lattice.

Under these conditions and operating the Globar at a brightness temperature of 1100°C, the calibration constant C for germanium is found to be 5×10^{15} cm^{-2}. If this technique is applied to study the semiconductor surface or to measure changes of the equilibrium carrier densities in the bulk as a function of position, Δp is then not equal to Δn and the absorption of coefficients for holes and electrons must be known separately.[4]

Sample preparation for any of the measurements discussed here is particularly simple. A sample of almost any size or shape will do. Basically all that is required is two parallel polished or etched flat surfaces through which the infrared beam may be passed. An uncut ingot may thus be examined longitudinally or transversely with only two prepared surfaces. Some of the measurements require ohmic while others require extracting contacts to the ends of the sample. The ohmic contacts for the present measurements were made by gold plating on a lapped surface.

In the infrared set-up used to date, the slit width was generally set at 0·01 cm. This distance is small compared to the diffusion lengths generally available for germanium and silicon nowadays. Thus, for measurements such as lifetime, a higher resolution is unnecessary. For thick samples the limit of resolution is set by the lack of collimation in the beam. The high index of refraction of semiconductors is always a help in constraining divergent beams to narrow paths. In our set-up the divergence of the beam outside of the germanium is 13°. For germanium (index of refraction of 4), the exit beam width for a sample 1-cm thick is then at least 0·05 cm. We improved the collimation, at a loss of infrared intensity, for some of our measurements. For thin samples or collimated beams the limit of resolution is set by diffraction effects. Therefore a beam a few microns wide can be used.

DETERMINATION OF SAMPLE TYPE

The sample type can be determined from examining the nature of the added carrier distributions in the presence of an applied electric field. Using the well-known result that a carrier packet moves in the same direction as the minority carriers and knowing the direction of the applied field, the sample type can be immediately determined. The carrier disturbances can be induced by the use of injecting or extracting contacts or by creating excess hole–electron pairs optically.

Fig. 3 shows examples of added carrier distributions. Without an applied field the distribution is governed by diffusion and the characteristic length is the diffusion length, given by $L = (D\tau)^{1/2}$. If a field is applied, the distribution is either field-aided, where, in a strong field the limiting characteristic length is the drift length and is given by $L = \mu^* E\tau$; or the distribution is field-opposed, where the limiting characteristic length is the field-contracted diffusion length and is given by $L = D/\mu^* E$. The field-aided and field-opposed distribution lengths are always, respectively, greater and less than a diffusion length.[3] In the examples shown, it is immediately evident that the sample is p-type since for an electric field pointing from right to left the distribution is field-aided, hence the electrons are the minority carriers. The active contact in Fig. 3(a) gave injection regardless of the direction of current flow[5], thus, the change in carrier distribution resulting from a reversal of the applied field could be demonstrated directly. The region to the left of the contact is field free, hence the carrier distributions are governed only by diffusion and recombination in both cases. The extraneous tail on the curve ii is caused by sample inhomogeneity and will be discussed later. Fig. 3(b) shows similar curves for the same sample where curves i and iii were obtained with the aid of n- and p^+-contacts, respectively, while curve ii corresponds to optical injection with no applied field. Similar curves have been obtained when optical injection was used throughout and the electric field was applied through ohmic (non-injecting) contacts. The effect of the applied electric field on the distribution length, as discussed above, is again

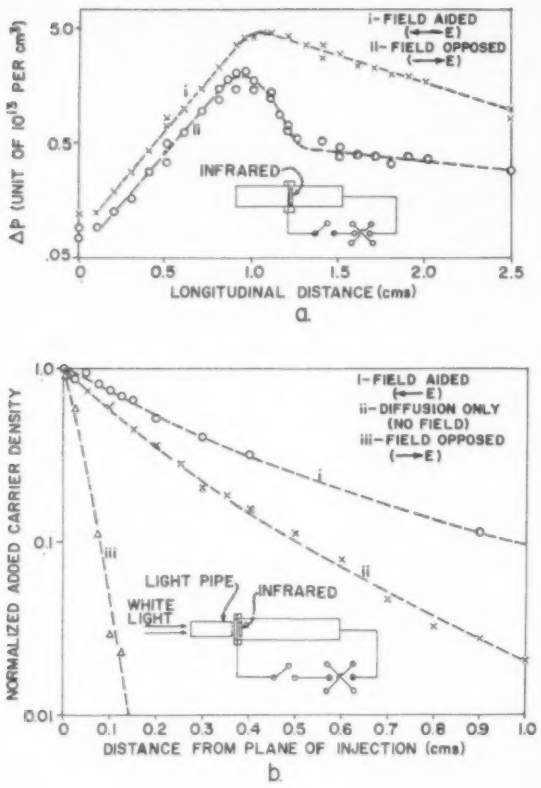


FIG. 3. Distributions of added carrier packets, determined by the use of infrared radiation, demonstrating the effect of an applied field. Such measurements are used to determine semiconductor sample type.

evident in Fig. 3(b). The optical-injection technique is simpler than the electrical-injection technique since it is easier to prepare ohmic contacts than to prepare injecting contacts; furthermore, the light for hole–electron pair creation can easily be piped to any point along the length of the sample.

A similar check of the sample type is, of course, automatically made in the drift-mobility experiment.[1] If the infrared apparatus is set up, a check of the type outlined here can be done with ease. Measurements such as these have enabled us to determine sample type in many instances when no conclusion could be reached from measurements of the sign of the thermoelectric power.

DETERMINATION OF SAMPLE-DOPING DENSITY AND ITS HOMOGENEITY

There are three ways in which infrared radiation can be used to determine semiconductor homogeneity and doping density. In each case the measurements yield an indication of the local free-carrier density and/or variations thereof. Each of the three techniques will be discussed and an indication of the resistivity or doping range in which the particular approach is most useful will be given.

(a) *Measurement of the local equilibrium carrier density*

One way in which the local doping can be measured is to determine the total transmitted infrared power from point to point. In this type of measurement no contacts to the sample are required. This technique has been used to measure the resistivity of the recrystallized region of an alloyed junction[6] where geometry and size are such that standard techniques cannot be used. For absolute measurements in highly doped semiconductors, a calibration for the degree of absorption of infrared radiation by the free carriers is necessary since impurity scattering becomes important in heavily doped materials. This technique has a number of disadvantages for high-resistivity samples. Firstly, because holes absorb infrared more strongly than electrons, there is a region of ambiguity just as there is in resistivity measurements for near-intrinsic n-type samples (see Fig. 1). Secondly, because the free-carrier absorption coefficient is low and because the reflectivity losses are high, any slight non-uniformity in the sample surface can lead to variations in reflection loss which can completely obscure resistivity changes. Thus, the semiconductor surfaces must be highly polished and they must be flat and parallel. Thirdly, there appears to be some absorption in the 2–5μ range which is not due to free carriers and is not as yet accounted for.

(b) *Measurement of the maximum permissible carrier-extraction level*

The equilibrium carrier densities in a semiconductor sample are governed by the relationship

$$(N_A+p_g)\,n_g = n_i^2$$

Here n_g is the minority-carrier density while p_g represents an equal density of majority carriers (n_g and p_g are the carrier densities resulting from thermal excitation of electrons from the valence to the conduction band). N_A is the density of holes due to doping. By the use of field-aided extracting sample leaving behind only N_A carriers per cm^3. This extraction level, which is the maximum permitted with space-charge neutrality, can be measured both longitudinally and transversely with the aid of infrared radiation. When n_g is measured, the local doping density can be deter-

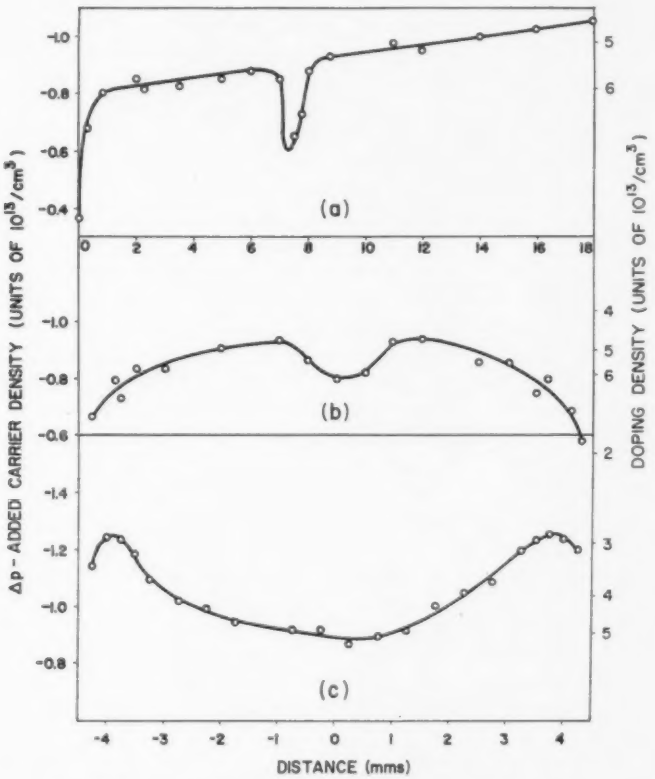


FIG. 4. Semiconductor doping density and homogeneity determined from an examination of a p-type sample under condition of complete extraction at 300°K. The extraction was achieved through the use of the p^+–p junction at position zero in curve (a). Curve (a) gives a measure of the longitudinal doping density while (b) and (c) represent the transverse doping density at two different positions. The doping density is related to the maximum extraction level with the aid of Fig. 1. The cross-sectional dimensions of the sample were 1 cm × 1 cm.

contacts (p^+–p and n^+–n contacts biased in the reverse direction[3]) and high sweeping fields, the drift length can be made greater than the length of the sample. Thus the thermally generated carriers can be almost completely extracted from the entire mined from the above formula or from Fig. 1. The sensitivity of this technique increases as the semiconductor doping density becomes less because the extraction level then increases. The temperature of the sample must be regulated and known

and near complete extraction must be achieved if absolute measurements are to be made.

Fig. 4 shows such measurements on a near-intrinsic, p-type germanium sample where information about both the longitudinal and transverse sample doping density and its homogeneity is obtained on one and the same sample when a current is passed through a p^+-contact in the reverse direction. The longitudinal measurements, Fig. 4(a), show a gradient in the doping, as might be expected in a pulled crystal, due to the increase in the concentration of the dopant in the melt as the germanium is used up. The extraction curve also shows that there is a more heavily doped region some 7 mm from the junction. It was not possible to detect this region of heavy doping from resistivity measurements. It was thus concluded that the heavily doped region was located deep within the bulk. The transverse measurement of Fig. 4(b) was taken in the vicinity of the heavily doped region of Fig. 4(a). These measurements also indicate that the region of heavy doping is located deep within the bulk. The decrease in extraction level near the edge of the sample is due to carrier generation at the surface[3] and not to an increase in doping density. Still further from the junction the transverse distribution of Fig. 4(c) is found. The extraction level here indicates that the crystal has a high-resistivity shell of 48 Ω-cm while the center has a resistivity of 43 Ω-cm. In such transverse measurements one always has to consider the carrier generation at the surface which tends to give a cosine distribution.[3] This effect of carrier generation at the surface can be decreased by the use of high sweeping fields.

The examination of a crystal under condition of complete extraction is a powerful approach for the determination of sample homogeneity and doping density, particularly for near intrinsic semiconductors, although it has been used successfully on 5 Ω-cm germanium. This technique can be used to reveal inhomogeneities in a distance much less than a diffusion length, depending on the width of the infrared beam. A special advantage of this method is the simplicity of the correlation of the experimental results with sample doping density. Furthermore, this technique is most sensitive for near-intrinsic samples where resistivity measurements are either insensitive or misleading. The sensitivity for any sample can be increased by working at elevated temperatures when n_q becomes larger.

(c) *Measurement of injection–extraction effects due to local inhomogeneities*

In the third method of detecting semiconductor inhomogeneities, a current is passed through the sample and any inhomogeneities are made to reveal themselves by causing local-injection or extraction effects. This approach to determine sample inhomogeneity of doping density is the converse of the technique described by Oroshnik and Many[2]; they add carriers to the sample and measure the resulting voltage while we apply a voltage to the sample and measure the local change in carrier density. To understand the mechanism involved one must recall that any changes in doping with distance can be considered as junctions which give rise to injection or extraction when a current passes through them. The contacts to the sample should be ohmic since any injection or extraction effects from the end contacts only complicate the interpretation of the results. This technique is analogous to that used in the investigation of metal-to-semiconductor contacts where a current is passed through the contact and, from the injection–extraction effects in the bulk, the nature of the surface and the contact can be ascertained.[5] The nature and degree of inhomogeneity can be determined in the same way from local measurements of injection and extraction as a function of current. If the patterns are complicated, translation from the injection–extraction curves into absolute inhomogeneity values is difficult. This technique is a powerful one for near-intrinsic samples where Fig. 1 shows that a change in doping may be accompanied by little or no change in sample resistivity.

Generally this technique will give only a qualitative measure of the inhomogeneity except for some special cases. Two such cases which give quantitative results can easily be considered. The first case is that of an abrupt junction in an otherwise homogeneous sample. Here junction theory may be applied directly to determine relative doping densities on either side of the junction. For example, for a heavily doped p–n junction, the injected-carrier densities and the doping densities on either side of the junction are related in the following way:

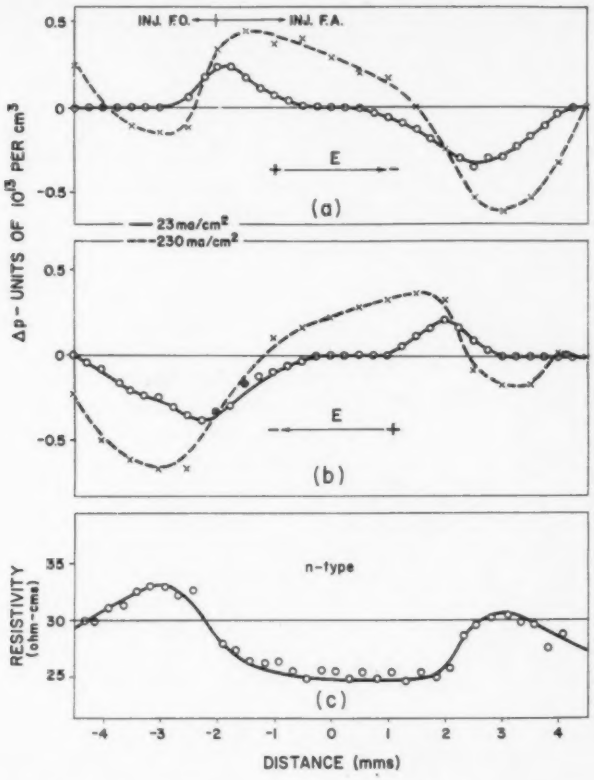


FIG. 5. Local injection and extraction effects resulting from the action of an electric field on a built-in carrier gradient. The qualitative correlation of the injection–extraction curves with the resistivity profile of curve (c) is immediately evident. These measurements give a measure of the transverse variations in doping density of a near-intrinsic, p-type ingot. F.A. and F.O. refer to field-aided and field-opposed distributions, respectively.

$$\frac{\Delta p_n}{\Delta p_p} = \frac{N_A}{N_D}$$

Similar expressions exist for light doping.[7] This case of abrupt junctions has been critically tested in the study of junctions in germanium.[7]

The second case is that of a gradual change in doping when it can be assumed that diffusion effects can be neglected with respect to drift effects and that the applied field is greater than any built-in field. If these conditions are fulfilled, the continuity equation for holes in a p-type sample is

$$\frac{dJ_p}{dx} = q\mu_p E\frac{dp}{dx} + q\mu_p p\frac{dE}{dx} = -q\frac{\Delta p}{\tau}$$

Using the well-known relationships for a p-type semiconductor

$$p = p_0 + \Delta p = N_A + p_g + \Delta p$$

$$p_g = n_g$$

$$n_g = n_0$$

and $n_0 p_0 = n_i^2$

the gradients of the equilibrium carrier densities are found to be related in the following way

$$\frac{dp_q}{dx} = \frac{dn_q}{dx} = -\frac{n_q}{p_0}\frac{dp_0}{dx}$$

$$= -\frac{dN_A}{dx}\left(\frac{n_q}{p_0} - \frac{n_q^2}{p_0^2} + \frac{n_q^3}{p_0^3} \cdots\right)$$

$$= -\frac{dN_A}{dx}\left(\frac{n_0}{n_0+p_0}\right)$$

and

$$\frac{dp_0}{dx} = \frac{dN_A}{dx}\frac{p_0}{n_0+p_0}$$

The above continuity equation thus becomes

$$\frac{dN_A}{dx}\left(\frac{p_0}{p_0+n_0}\right) + \frac{d\Delta p}{dx} + \frac{p}{E}\frac{dE}{dx} = -\frac{\Delta p}{\mu_p E\tau}$$

For near-intrinsic samples where dE/dx may be very small because the resistivity is very insensitive to doping density, the gradient of the doping density is given by

$$\frac{dN_A}{dx} = -\frac{p_0+n_0}{p_0}\left[\frac{\Delta p}{\mu_p E\tau} + \frac{d\Delta p}{dx}\right]$$

Other approximations might be made for samples where the doping density is higher. A similar expression can be derived for n-type materials.

If this relationship is applied to calculate the gradient in the doping density which gives rise to the tail on the curves ii of Fig. 3(a), a change in doping density of about 10 per cent per cm is found. This is in agreement with observations made from extraction measurements. This technique was used to trace down sample inhomogeneities in some earlier work.[3]

Fig. 5 gives an example of a test of the transverse homogeneity of an ingot using this technique and, for comparison, standard resistivity measurements on the same sample. The measurements were made on a bar of dimensions 1·8 cm × 0·4 cm × 0·6 cm where the length of the bar corresponds to a transverse direction in the original ingot. The qualitative correlation of the injection–extraction effects with that expected from the resistivity profile of Fig. 5(c) is immediately evident; viz. for the n^+–n junction, injection and extraction in the forward and reverse directions, respectively.

If the nature of the carrier distributions in Fig. 5 is taken into consideration, the sample type can also be determined. This can be done as follows. By first noting the symmetry of the low-current Δp versus distance measurements, it can be assumed that the sample resistivity is roughly symmetrical about the center of the ingot with fairly abrupt changes in resistivity occurring at +0·2 cm and −0·2 cm. By next noting that injection occurs at −0·2 cm and that the direction of the current is from left to right, it can be concluded that the nature of the junction must be p–n, p^+–p or n–n^+. From an examination of the higher current-injection profile, the injected-carrier distribution is seen to be field-opposed on the left of −0·2 and field-aided on the right. These carrier distributions and electric field directions can apply only to an n–n^+ type of structure (see Fig. 2 Ref. 7). This conclusion is consistent with the type and resistivity measurements determined from standard techniques, shown in Fig. 5(c). From an examination of the carrier distributions in other parts of the sample, additional details regarding the homogeneity can similarly be obtained.

As a further demonstration of this technique, two additional examples on the transverse non-uniformity of pulled crystals are provided in Figs. 6 and 7. Fig. 6 demonstrates the situation where resistivity measurements indicate that the crystal is highly uniform yet the infrared technique reveals significant non-uniformity. The crystal was first determined to be p-type with the aid of the infrared technique. From analysis of the injection–extraction curve, as discussed in the last paragraph, it was shown that the ingot must have a p^+-shell. This conclusion is further confirmed with the aid of Fig. 6(c) where it may be seen that the activity for one direction of current flow collapses almost entirely when 2 mm of the sample is removed from one end. The 2 mm should thus be a measure of the thickness of the p^+-shell.

Fig. 7 shows measurements on a 6 Ω-cm n-type sample. Because of the n^+–n–n^+ structure, the extraction in the n-region will be field-aided while the injection will be field-opposed, hence field-aided extraction will predominate at high currents regardless of the direction of current flow. Thus, as discussed in sub-section (b), the local sample doping can be determined from the measured

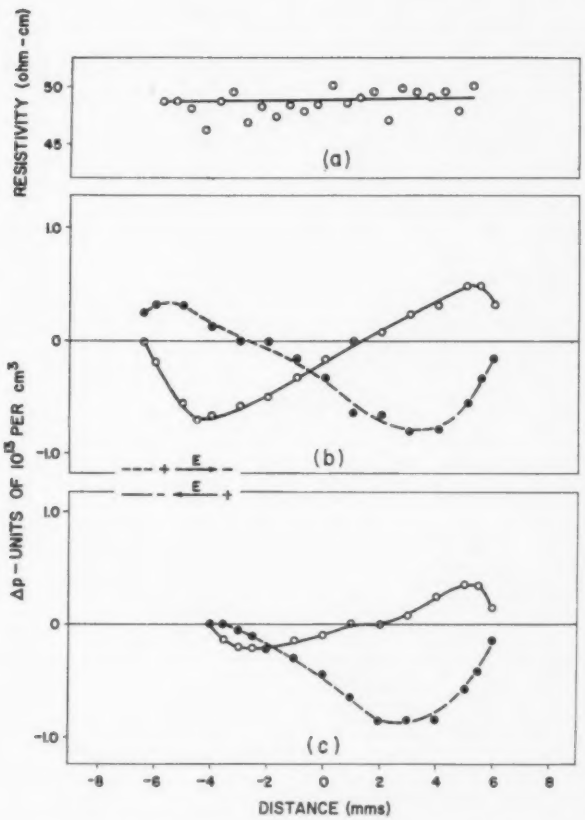


FIG. 6. Injection and extraction effects occurring in a sample which revealed no variations in resistivity. The curve (c) shows how the activity collapses at one end of the sample when 2 mm of the sample is cut away.

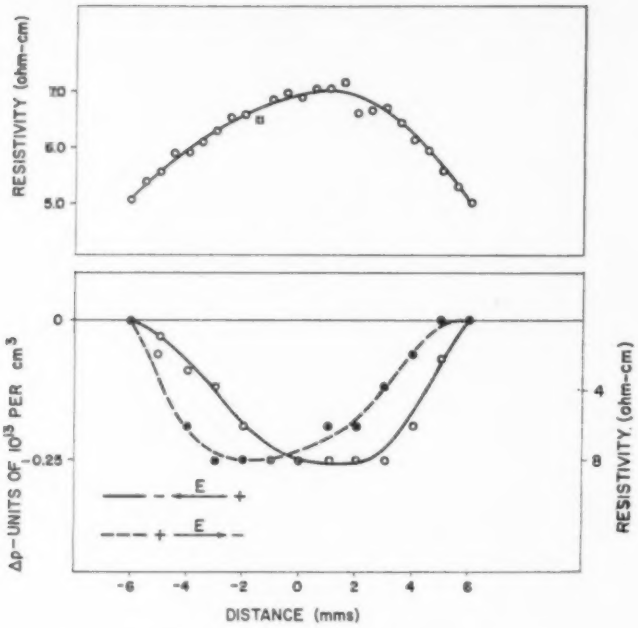


Fig. 7. Extraction effects in a 6 Ω-cm, *n*-type sample. The correlation of the extraction level with resistivity measurements is again evident.

extraction level with the aid of Fig. 1. It can be seen from Fig. 7 that the local sample-doping density determined in this way agrees with that determined from resistivity measurements except near the contacts where the extraction level is too low because of the high generation rate at the contacts. A similar predominance of extraction at higher currents was also found for the sample in Fig. 6 which has a p^+–p–p^+ structure.

As implied earlier, the conversion of the injection–extraction curves to inhomogeneities can be complicated. Variations in doping density which occur within a diffusion length cannot be resolved although their presence can be inferred. If the structure happens to be one of those found in the examples of Fig. 6 or 7 where field-aided extraction will prevail at high currents, the sample local-doping density can be determined easily with the aid of Fig. 1, as discussed in sub-section (b). When this is not the case, however, the injection–extraction effects are still useful for tracing down local inhomogeneities, as shown in Fig. 5.

Measurements have also been made on uniform low-resistivity crystals. As might be expected, little or no activity was found even at high currents for these crystals.

Acknowledgements—The author is indebted to R. C. HUGHES and P. C. MURAU for the use of their numerous crystals and resistivity measurements from which interesting samples were selected to demonstrate the power of the infrared technique, particularly in the measurements described in sub-section (c). He is also indebted to Dr. F. K. DU PRÉ for helpful discussions and to Mr. H. W. KEESE for technical assistance.

REFERENCES

1. N. J. HARRICK, *J. Appl. Phys.* **30,** 451 (1959).
2. J. OROSHNIK and A. MANY, *Solid-State Electron.* **1,** 46 (1960).
3. N. J. HARRICK, *Phys. Rev.* **103,** 1173(1956).
4. N. J. HARRICK, *J. Phys. Chem. Solids* **14,** 60 (1960).
5. N. J. HARRICK, *Phys. Rev.* **115,** 876 (1959).
6. F. E. ROBERTS, *Solid-State Electron.* **1,** 93 (1960); also private communication, to be published.
7. N. J. HARRICK, *J. Appl. Phys.* **29,** 764 (1958).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 245–249. Printed in Great Britain

NOTES

A continuous-reading four-point resistivity probe

(*Received* 5 *April* 1960; *in revised form* 19 *May* 1960)

THE four-probe resistivity-measurement technique provides a simple non-destructive method for the measurement of resistivities up to 100 Ω-cm. The main difficulty with the techniques used so far is that each measurement has involved a lifting and lowering of the probe head because the probes have been pointed tungsten wires. This process has been made automatic[1] but it still provides only a series of point measurements.

The present apparatus uses ball-point probes and can be run along the surface of the specimen providing a continuous signal from which a resistivity profile may be derived.

The probes are the ball ends of "Biro" pens. These consist of a metal ball about 0·5 mm diameter held in a metal tube. This tube was turned down to 2 mm diameter to allow small probe spacings. The probes were a sliding fit in a $\frac{1}{4}$-in. thick Perspex block and were spaced 2·5 mm between centres. In use this spacing has not changed appreciably over several months. The probes are retained in their holes by wire-leaf springs which also provide the electrical connexions.

In use the head was fastened to an arm driven at a constant speed by a lead screw and motor system. To provide the necessary contact pressure a weight was placed on the arm. For germanium this provided a force of about 800 g on to the specimen. For softer materials less would be required.

A constant current is fed between the two outer probes from the anode circuit of a battery-supplied pentode valve and the e.m.f. between the inner probes is displayed on a 10-mV full-scale potentiometric recorder.

The results are calculated from the formula

$$\rho = \frac{V}{I}\,\frac{2\pi}{1/S_1+1/S_3-1/(S_1+S_2)-1/(S_1+S_3)}$$

where ρ is the resistivity, V the e.m.f. between inner probes, I the current between outer probes and S_1, S_2, S_3 are the probe spacings. This assumes the region under the probes to be homogeneous.[2]

In order to eliminate the effect of minority-carrier injection from the probes, the surface of the sample was lapped with 800-mesh grinding powder before measurement.

Over a number of runs the results are reproducible to about 1 per cent and are consistent with the measurements of a conventional four-probe head to within about 2 per cent.

J. C. BRICE
A. A. STRIDE

Mullard Research Laboratories
Cross Oak Lane
Salfords, Nr. Redhill
Surrey.

REFERENCES

1. D. DEW HUGHES, A. H. JONES and G. E. BROCK, *Rev. Sci. Instrum.* **30,** 920 (1959).
2. E. B. HANSEN, *Appl. Sci. Res., Hague* **B8,** 93 (1960).

Creep and forming in selenium rectifiers

(*Received* 14 *June* 1960)

AN EXAMINATION of more than a decade of literature and of the most recent advanced text in the subject[1–5] shows that no adequate explanation has appeared for certain time-dependent changes, commonly called "creep" or forming effects, in the rectifying junction of the selenium rectifier. It is in general agreed that ion diffusion takes place, especially under the influence of applied fields, and that a layer of cadmium selenide grows on the surface of the selenium,[6, 7] but these processes are quite inadequate to explain the variety of effects. Although occasional reference has been made to the fact that the distribution of reverse-current density over the rectifier surface is highly nonuniform,[2, 8] almost all previous workers make the assumption that explanations are to be

sought in changes occurring over the whole area of the rectifier. "Positive creep" was at one time ascribed to temperature effects,[1] but later this interpretation appeared to be possible but not essential.[3] It is the purpose of this note to show that the solution to the problem lies predominantly in the unique behavior of small regions where the rectifying junction is weak, and where large deviations in both current density and temperature make ordinary junction analysis quite irrelevant. TOMURA[2] comes closest to our present findings when he shows that under a transparent counter-electrode local color spots develop during electrical forming, almost certainly due to the formation of cadmium selenide between the counter-electrode and the selenium. To explain the reduction in reverse current during forming, he suggests "Joule's heat produced by heavy currents through the local spots during forming* may melt the selenium semiconductor there, changing it to another modification which is an insulator and diminishing the effective area of the spots through which current can flow". Apart from this isolated comment on the melting of selenium, TOMURA's ideas on rectifier behavior do not correspond to the present analysis.

The model which we have successfully used to explain almost all published phenomena (apart from the very long-term selenide growth effects), and our new observations, is as follows.† It is assumed that under reverse-voltage conditions, indefinitely large current densities can exist at local weak spots in the rectifying junction, but that the imperfection can be extremely small, thus limiting the actual magnitude of the current. The result is that in regions possibly ranging down to microns in size, temperatures may easily reach the melting point of selenium,‡ even for currents below 1mA. Since molten or vitreous selenium is an insulator, its production at the site of the imperfection results in a protective mechanism which has the effect of controlling the current

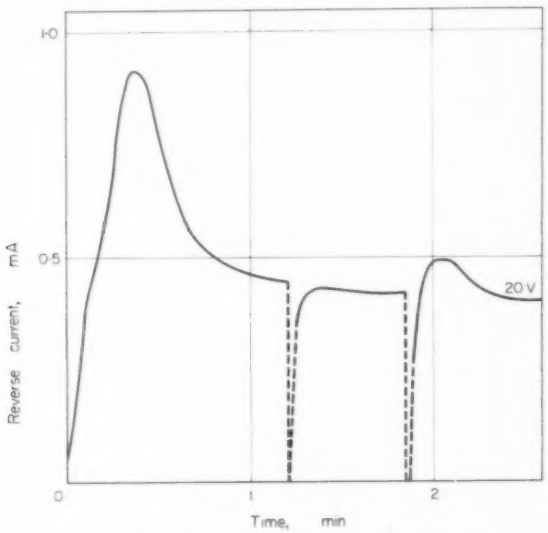


FIG. 1. Current creep in an unformed selenium rectifier with 20 V reverse voltage applied. Voltage is switched off for 1-sec intervals in the negative-creep region.

* A slight rewording of the English translation has been given.

† The model is derived from some high-temperature studies in selenium rectifiers.[9]

‡ Note that the counter-electrode was a 140°C m.p. alloy of 60 per cent Bi, 40 per cent Cd. We frequently melted parts of the counter-electrode during experiments without adverse effect. This in essence refutes a claim by BILLIG[10] on breakdown criteria.

density and halting what would otherwise be a destructive runaway condition. The model is incomplete without some further specifications. Under current-carrying conditions, the system must be reversible, that is, a dynamic equilibrium exists which adjusts the amount of molten selenium and fixes the local current density according to the heat dissipation from the hot center to the cooler surroundings. If the rectifier is cooled, part of the molten selenium freezes, increasing the local current, which provides the necessary increase of heat generation to maintain a molten center in a cooler environment.

A typical classical current–time graph, Fig. 1, showing positive and negative creep and switch-off effects* may now be readily interpreted. The initial rising part of the plot reflects the rapidly rising temperature of the weak spots in the junction and their deterioration in electrical characteristics. The falling part of the curve, or the "negative creep" region, is attributable to the melting of selenium at the hot spots, to the degree necessary for the final steady-state condition. If the current is switched off momentarily, a new current level is established below that before switch-off, and positive creep takes place toward the old value. The physical explanation is that during the switch-off time, the hot spots cool significantly, become more resistive, and carry a reduced current upon switching on. The rewarming of the hot spots results in renewed positive creep to re-establish the steady state. Some of the overshoot of the current after switch-off must be attributed to the recrystallization of vitreous selenium, since the region drops in temperature, so that the current must to some degree pass through the original cycle when the rectifier was first subjected to voltage.

We present at this time two critical experiments to support the model. Fig. 2 shows, superimposed upon a normal creep curve, current–time traces which are obtained at various points in the cycle if the rectifier is air-cooled with a blower. On the rising part of the curve, the rectifier behaves in "normal" fashion and the reverse current drops upon cooling. Sufficiently far into the negative-creep region, an "inverse cooling effect" appears where the current rises to high values upon cooling. It is possible to demonstrate a wide variety of combinations of the two effects with variations in rectifier history, voltage and ambient temperature. The interpretation is based on the absence of molten selenium on the rising part of the creep curve, where cooling the rectifier simply reduces the temperature of hot crystalline selenium in the barrier region very close to the weak spots, and on the presence of molten selenium in the negative-creep section where cooling, if not too intense, results in partial recrystallization of the molten

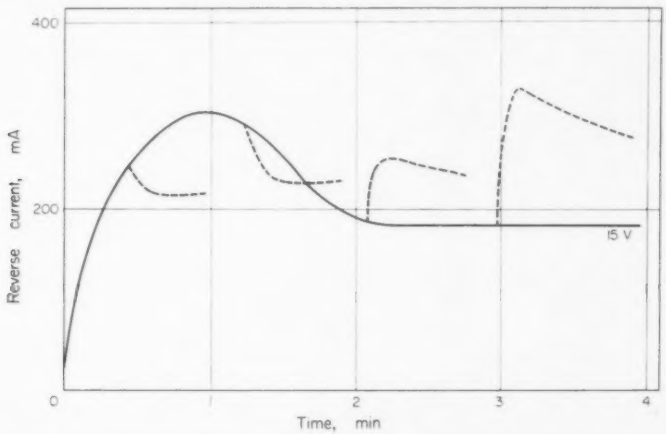


FIG. 2. Cooling effects superimposed at various points on the current–creep curve of an unformed selenium rectifier with 15 V reverse voltage applied. The current follows the dashed line if the rectifier is air cooled at the indicated points.

* See Fig. 9 of Ref. 1 or Fig. 41.4a of Ref. 5.

selenium and a better conduction path through hot selenium which is closer to the original central defect. If the cooling is sufficiently strong the current may be brought down from the high value as the hot spots crystallize completely and all of the selenium cools down.

The second experiment is represented in Fig. 3 where a rectifier which has come to an approximate steady state in the negative-creep region is subjected to small step changes in applied voltage. For a small change downwards, the current drops instantly to a lower value and then creeps upwards essentially to its original value. For a step-up in voltage, the current steps up and then decays close to the original value. It will be observed that the time constant for the decay after a step-up in voltage is significantly shorter than for the case of a step-down in voltage.

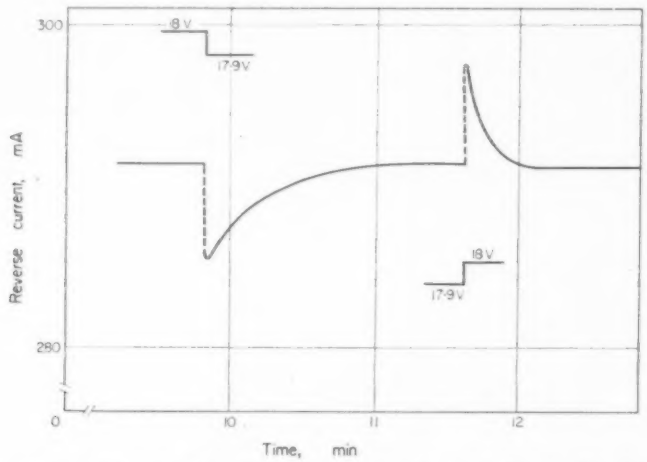


FIG. 3. Transient changes in reverse current caused by step changes in reverse voltage on a selenium rectifier after approximate equilibrium is established in the negative-creep region.

The phenomenon is essentially identical with earlier observations[9] and is explained on the basis that a drop in voltage reduces the power dissipated in the hot spots. The resulting slight drop in temperature below the freezing point of selenium leads to recrystallization upon which the current rises until a new equilibrium is established. A rise in voltage results in melting of selenium and a fall in current. The time-constant effect for the current recovery is explained by the higher rate at which selenium melts compared with its crystallization rate under the experimental conditions. The fact, that in this experiment the system reacts to voltage changes in such a way as to hold almost constant the current–voltage product, comes as close to a proof of the model, in view of the earlier work,[9] as one could wish.

In more detailed publications we plan to show that essentially all of the obscure phenomena previously reported for the selenium rectifier in the categories of creep, forming, breakdown, resistance instability and, to some extent, aging, may be explained on the basis of the above model.

Acknowledgement—One of us (R. W. T.) would like to acknowledge the financial assistance of the Office of Naval Research during this work.

A. C. ENGLISH
R. W. THOMAS

Department of Electrical Engineering
University of California
Berkeley 4,
California

REFERENCES

1. H. K. HENISCH and J. EWELS, *Proc. Phys. Soc. Lond.* B**63,** 861 (1950).
2. M. TOMURA, *J. Phys. Soc. Japan* **5,** 222 (1950).
3. R. COOPER and J. HARRINGTON, *Proc. Phys. Soc. Lond.* B**65,** 303 (1952).
4. R. COOPER, *Proc. Phys. Soc. Lond.* B**65,** 409 (1950).
5. H. K. HENISCH, *Rectifying Semiconductor Contacts.* Clarendon Press, Oxford (1957).
6. S. POGANSKI, *Z. Elektrochem.* **56,** 193 (1952).
7. M. TOMURA, *Bull. Chem. Soc. Japan* **22,** 82 (1949).
8. K. LEHOVEC, *Z. Phys.* **125,** 451 (1949).
9. A. C. ENGLISH and W. H. TOBIN, *Commun. Electron.* **79,** 9 (1960).
10. E. BILLIG, *Proc. Roy. Soc.* A**207,** 156 (1951). See especially p. 170.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 250–251. Printed in Great Britain

BOOK REVIEWS

A. K. JONSCHER: **Principles of Semiconductor Device Operation.** Bell, London, 1960. 168 pp., 30*s*.

NUMEROUS books on semiconductors and transistors, covering a wide range of academic level, have been published in recent years. Dr. JONSCHER points out in his preface that these are either on semiconductor physics with a brief and inadequate chapter on devices at the end, or on devices from the electronic engineer's angle with a brief and inadequate chapter on physics at the beginning. The present book lies between these two extremes and fills a very real gap. The more subtle, but nevertheless important, properties of transistors and similar devices are by no means self-evident from a discussion of basic principles. Dr. JONSCHER has shunned the grossly over-simplified treatments which are used so often and, if he is not always rigorous, at least he points out any approximations used with refreshing honesty.

The subject matter is limited to devices relying upon carrier-density disturbance, e.g. transistors of all kinds, *p–n–p–n* diodes, etc. The approach is essentially theoretical and the standard is that of a postgraduate course to serious students of the subject. There is no mention of device-fabrication techniques (except where these influence the mathematical assumptions that may be made) and the book might easily have been called "Theory of . . ." rather than "Principles of . . .". The quality of exposition is, in my opinion, first rate: I *enjoyed* reading it. The information rate per page is high, but not indigestible, and the collected list of symbols at the back very valuable.

An interesting feature of the text is the problems set at the end of each of the six chapters. The problems form an extension, rather than a revision, of the text. With one or two exceptions they are extremely well thought out and should not be skipped over, though I see little point in a solitary reader tackling a question on Chapter I which begins "Discuss . . .". The appendices, of which there are eight, are properly used to avoid discontinuities in the main text, which leads the reader gently, but firmly, to the final chapter on carrier flow in inhomogeneous media. References are given at the end of each chapter. Wherever possible these are to other textbooks and review articles rather than original papers. There is a comprehensive index.

This book should be of interest to Physicist and Device Engineer alike. It can be recommended to anyone learning or teaching transistor physics as comprehensible and accurate. My only regret is that it is so short. A companion volume, from the same source, covering the physics of other semiconductor devices such as Tunnel Diodes, Field-Effect Devices, etc., would be most welcome.

ALAN F. GIBSON

D. A. BELL: **Electrical Noise. Fundamentals and Physical Mechanism.** Van Nostrand, London, 1960. x + 342 pp., 50*s*.

THE lack of a book providing comprehensive up-to-date information and references on the mechanism of noise in electric circuits and solid-state devices was widely felt, thus the appearance of the book under review will be warmly welcomed, and even more so since the author is an expert who has carried out much original research in the field and can write clearly and attractively.

The originality is amply reflected in the book, so that it may be that other experts in various aspects of the field will find that the author's views and approach have been given undue prominence. It must be said though that parallel contributions are fairly reviewed, and it is of course a great advantage to have such a vast field covered from a single point of view. To more modest research workers the book will be invaluable as a guide to everything of importance discovered on the subject up to 1959. On some points, such as on noise in metal films, the information was particularly scattered.

After an interesting historical introduction and an exposition (also original in some aspects) of mathematical techniques and fundamental physical

principles, Johnson noise and shot noise are considered in detail, but in the framework of a unified theory. The section on u.h.f. valves covers much recent work on various devices, though, perhaps inevitably, the treatment (e.g. of masers) is not as exhaustive as in previous chapters.

Of particular interest is the chapter on the $1/f$ spectrum of current noise in semiconductors. In these last few years it has become clear that the phenomenon is due to a fluctuation in the number of carriers in the conduction band, while surface effects, traps, etc., play a very minor part, if any. The author believes that the inverse frequency spectrum must be due to some form of co-operative phenomenon and that the problem is in fact that of the waiting time in a queue which is served by a number of outlets. The theory is not yet completely conclusive, but is already infinitely more convincing than previous theories postulating special time constants or trapping mechanisms.

The concluding chapters deal with radiation detectors and standards and measuring techniques. Another chapter gives a synthetic review of Barkhausen noise in ferromagnetics. Specialized mathematical points are given in eighteen appendices. There is a Subject Index, but it is too limited to be of much use. Although quite lucid, the book requires careful and thoughtful reading, and then it will be found most illuminating and stimulating.

The volume is well produced and can be strongly recommended both to physicists—and not only to those interested in the applied side of the subject—and to electrical engineers. It should be useful both at an advanced level and as an introduction to the subject.

L. PINCHERLE

Solid-State Electronics Pergamon Press 1960. Vol. 1, p. 252. Printed in Great Britain

ERRATUM

M. GREEN and I. N. GREENBERG: Transient charge-carrier diffusion in high-purity silicon. Vol. 1, No. 1, pp. 85–92. Half-tone plates between pp. 88 and 89. Fig. 6 was placed over the caption for Fig. 5, and Fig. 5 over that for Fig. 6.

EDITORS' FOREWORD

THIS special issue of *Solid-State Electronics* brings to our readers twenty-three of the papers presented at the Symposium on Superconductive Techniques for Computing Systems sponsored by the Information Systems Branch of the Office of Naval Research and held in Washington, D.C., on 17, 18 and 19 May 1960.

At this Symposium an attempt was made to define the probable role of superconductive devices in the computer and information processing field. Although some attention was given to a comparison of superconductive devices with other techniques of handling problems in information processing, it developed that there remain serious difficulties in the use of superconductive cells in logic circuits, and the main business of the Symposium revolved around the problems of reducing superconductive logic and memory devices to a really practical state. At any rate, the Editors feel that the material is of very real interest to all workers in the field of solid-state electronics and congratulate the Conference Committee on the quality of papers presented at the Symposium.

The Editors are highly indebted to the Symposium Committee and especially to its Chairman, Dr. MARSHALL C. YOVITS, for their cooperation in the arrangements for this publication. Other members of the Symposium Committee include RICHARD S. COLLIER (Department of Defense), CARL G. GARDNER (Department of Defense), Commander VANCE R. WANNER (Office of Naval Research) and RONALD L. WIGINGTON (Department of Defense).

Because of other commitments, several of the papers presented at the Symposium are not included in the present collection. A more complete Proceedings has been published by the Office of Technical Services, Department of Commerce, under the sponsorship of the Office of Naval Research.

BLANK
PAGE

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 255–257. Printed in Great Britain

OUTLINE OF RECENT DEVELOPMENTS IN SUPERCONDUCTIVITY*

BERNARD SERIN

Rutgers University, New Brunswick, N. J.

RECENT experimental results in superconductivity are distinguished for having elucidated the microscopic structure of the superconducting phase. They have also raised important new questions.

It has proved most useful to describe many of the properties of superconductors in terms of a gap in the one-electron energy spectrum[1] which is temperature dependent. The BARDEEN *et al.*[2] (BCS) theory makes explicit predictions concerning this energy gap, as well as other properties of superconductors. The BCS result for the density of electronic states as a function of energy is shown in Fig. 1. E_F is the Fermi energy and 2ϵ is the energy gap. To obtain the number of

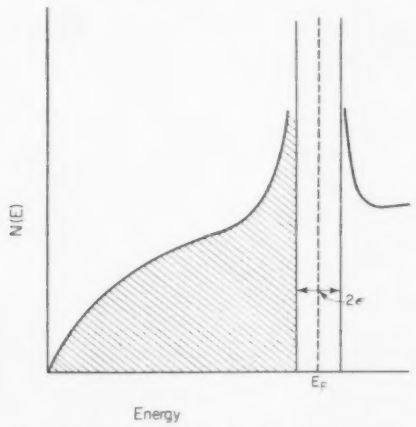


FIG. 1. The density of states per unit-energy range, $N(E)$, as given by the BCS theory, is shown as a function of the energy.

electrons of a certain energy, this curve is multiplied by the Fermi–Dirac distribution function. As a result, at 0°K all the states of energy less than $(E_F-\epsilon)$ are filled with electrons as shown by the cross-hatching in Fig. 1, and no states of energy greater than $(E_F+\epsilon)$ are occupied. Under these circumstances all the electrons have superfluid properties. As the temperature is raised, some electrons are thermally excited across the gap as represented by the high-energy tail of the distribution function. The electrons having energies on the high side of the gap behave in the normal manner. Thus, the BCS theory can be expressed in terms of a two-fluid model of the superconductive phase. Also, with increasing temperature the size of the energy gap decreases as shown in Fig. 2, until the gap vanishes at the transition temperature and the density-of-states function reverts discontinuously to a usual one.

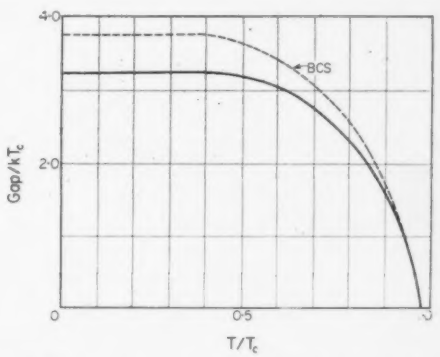


FIG. 2. The energy gap is plotted as a function of reduced temperature. The solid curve is experimental[4], and the dashed curve is the result of the BCS theory. T_c is the transition temperature, and k is Boltzmann's constant.

Direct experimental evidence for the existence of the energy gap has been provided by the observations of absorption of quanta of electromagnetic radiation by superconductors.[3,4] RICHARDS and

* This work was supported by the National Science Foundation.

TINKHAM,[3] using infrared techniques, sent radiation of wavelength 0·15–2 mm into a superconductor in the form of a black body and determined the energy absorbed as compared with the absorption when the cavity was in the normal phase. At a given temperature, they observed that additional absorption suddenly appeared in the superconductive phase when the frequency of the radiation was greater than a critical value corresponding to the quantum energy necessary to excite electrons across the gap. BIONDI and GARFUNKEL,[4] using microwave techniques, determined the surface resistance as a function of temperature of an aluminum wave-guide for wavelengths 3–19 mm. As in the foregoing investigation, at a given temperature additional resistance appeared when the frequency exceeded a critical value. The results obtained[4] for the dependence of energy gap on temperature are shown in Fig. 2; the agreement with BCS is good.

Measurements of the attenuation of longitudinal ultrasonic waves[5] also reveal the dependence of gap size on temperature. According to BCS, the ratio of the attenuation in the superconductive phase to the attenuation in the normal state is $2[\exp(\epsilon/kt)+1]^{-1}$. Thus, measurements of this ratio yield ϵ as a function of temperature. Agreement between theory and experiment is again good.

It is implicitly assumed by BCS that the Fermi surface and the energy gap are isotropic in momentum space. There have, however, been recent suggestions that the gap may be anisotropic. COOPER[6] has suggested that specific-heat data of the kind illustrated in Fig. 3 might be explained by anisotropy in the energy gap. As shown schematically in Fig. 3, the electronic specific heat of the superconductive phase deviates at low temperatures from the high-temperature exponential dependence on reciprocal temperature. Two gap energies, one large and one small, could give a curve of the observed shape. Also, recent measurements[7] of ultrasonic attenuation in single crystals, seem to yield different values for the gap energy at 0°K when the sound is propagated along different crystal directions. Data on nuclear-spin relaxation[8] can perhaps be given a similar interpretation. Another interesting question has been raised by the observation of apparent structure in the energy gaps of lead and mercury.[9]

Although time does not permit detailed discussion, it seems appropriate to mention that interesting measurements and speculations have been made on the Knight shift[10], penetration depth[11], impurity effects[12], and ferromagnetic superconductors[13].

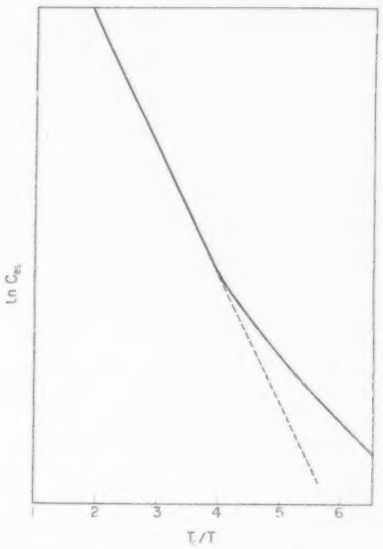


FIG. 3. The superconductive specific heat, C_{es}, is shown schematically as a function of reciprocal reduced temperature. The solid line is experimentally observed in several metals; the dashed line is expected from an isotropic energy gap.

REFERENCES

1. For a review up to 1958 of data pertaining to the energy gap see M. A. BIONDI, A. T. FORRESTER, M. P. GARFUNKEL and C. B. SATTERTHWAITE, *Rev. Mod. Phys.* **30,** 1109 (1958).
2. J. BARDEEN, L. N. COOPER and J. R. SCHRIEFFER, *Phys. Rev.* **108,** 1175 (1957). See also L. N. COOPER, *Amer. J. Phys.* **28,** 91 (1960).
3. P. L. RICHARDS and M. TINKHAM, *Phys. Rev. Letters* **1,** 318 (1958).
4. M. A. BIONDI and M. P. GARFUNKEL, *Phys. Rev.* **116,** 853 (1959).
5. See R. W. MORSE, *Progress in Cryogenics* Vol. *I*, pp. 219–259 (Edited by K. MENDELSSOHN), Academic Press, New York (1959).
6. L. N. COOPER, *Phys. Rev. Letters* **3,** 17 (1959); for a recent review of experimental specific-heat data see H. A. BOORSE, *Phys. Rev. Letters* **2,** 391 (1959).

7. R. W. Morse, T. Olsen and J. D. Gavenda, *Phys. Rev. Letters* **3,** 17 (1959).
8. A. G. Redfield, *Phys. Rev. Letters* **3,** 85 (1959).
9. D. M. Ginsberg, P. L. Richards and M. Tinkham, *Phys. Rev. Letters* **3,** 337 (1959).
10. F. Reif, *Phys. Rev.* **102,** 1417 (1956); **106,** 208 (1957); G. M. Androes and W. D. Knight, *Phys. Rev. Letters* **2,** 386 (1959); for recent theoretical interpretations see R. A. Ferrell, *Ibid.* **3,** 262 (1959); P. C. Martin and L. P. Kadanoff, *Ibid.* **3,** 322 (1959); J. R. Schrieffer, *Ibid.* **3,** 323 (1959); P. W. Anderson, *Ibid.* **3,** 325 (1959).
11. A. L. Schawlow and G. E. Devlin, *Phys. Rev.* **113,** 120 (1959); M. S. Dresselhaus and G. Dresselhaus, *Phys. Rev. Letters* **4,** 401 (1960); M. P. Sarachik, R. L. Garwin and E. Erlbach, *Ibid.* **4,** 52 (1960).
12. G. Chanin, E. A. Lynton and B. Serin, *Phys. Rev.* **114,** 719 (1959); for a theoretical interpretation see P. W. Anderson, *J. Chem. Phys. Solids* **11,** 18 (1959).
13. B. T. Matthias, *Proceedings of the Symposium on Superconductive Techniques for Computing Systems, Washington,* 1960. Office of Naval Research (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 258–260. Printed in Great Britain

THE USE OF SUPERCONDUCTIVE DEVICES IN RESEARCH AT LOW TEMPERATURES

I. M. TEMPLETON

National Research Council, Ottawa, Canada

THE thermoelectric power of a metal at low temperatures is generally of the order of $10^{-8}\ T$ V/deg., where T is the absolute temperature. For work in this field, at temperatures of about 1°K[1–4] one must therefore be able to measure with reasonable accuracy e.m.f.'s of about 10^{-8} V, i.e. one must be able to detect e.m.f.'s smaller than 10^{-9} V.

Our present system uses as detector a feedback galvanometer amplifier[5] in which the r.m.s. noise level is about 2×10^{-10} V. This is, in fact, the room temperature thermal noise from the input wiring and primary galvanometer (of about 5 Ω total resistance). Discrimination against thermal e.m.f.'s in the leads from liquid helium to room temperatures is achieved by using a superconducting reversing switch[6], as in Figs. 1 and 2, in the liquid helium bath. In this device the pairs of tantalum coils *aa'* and *bb'* are alternately switched by magnetic fields into the resistive condition, reversing the e.m.f. seen by the external detector.

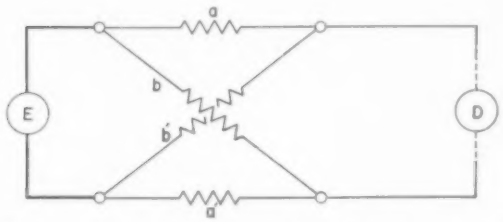


FIG. 1. Superconducting reversing switch and associated circuits. E is the source of e.m.f. to be measured and D is an external detector, e.g. potentiometer, or galvanometer amplifier. The heavy lines represent superconductive wiring.

If measurements of greater accuracy, or at lower temperatures, are required, this system will no longer be satisfactory. A substantial reduction of thermal noise may be achieved by operating the detector in liquid helium. The change of temperature alone will reduce the r.m.s. noise level by a factor of 10, and by using superconductive wiring, the primary circuit resistance may be reduced to that of the specimen plus that, if any, of the detecting device. Even if the total circuit resistance were as (comparatively) high as 10^{-4} Ω, the r.m.s. noise level would be reduced to 10^{-13} V.

A detecting device capable of operating in a liquid helium bath has been described by the author[7] (Figs. 3–5). In this case a modulated magnetic field causes a section of fine tantalum wire to act as a chopper by switching its resistance between zero and about 0·1 Ω. The tantalum is connected in series with the primary (one turn of heavy annealed copper) of a miniature transformer, so that it modulates the e.m.f. from the low resistance source. The limiting noise in this case is from the amplifier, and is equivalent to about 4×10^{-12} V at the low level transformer primary. In practice, it is necessary to interrupt

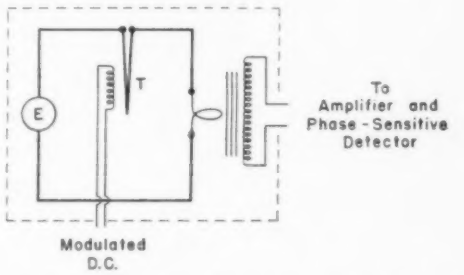


FIG. 3. Superconducting modulator. The e.m.f. E must be reversed or interrupted to identify the signal in the presence of spurious pickup from the modulator coil.

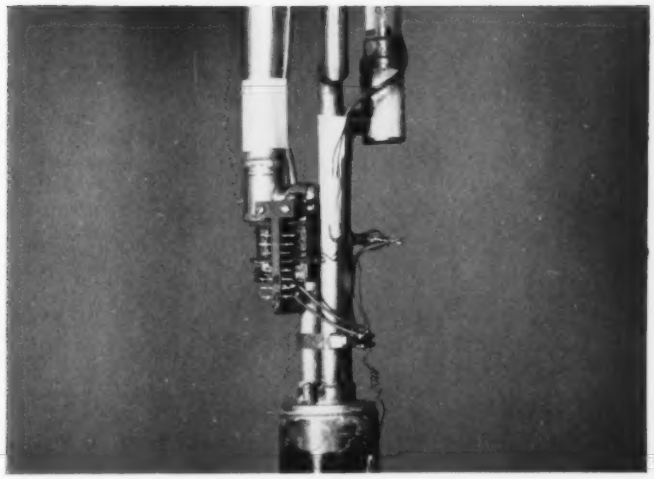


FIG. 2. Superconducting reversing switch in position on a cryostat. The lead (Pb) wires from the specimen are soldered into platinum tubes sealed into a soft glass bulb; approximately $\frac{2}{3}$ natural size.

[*facing p.* 258

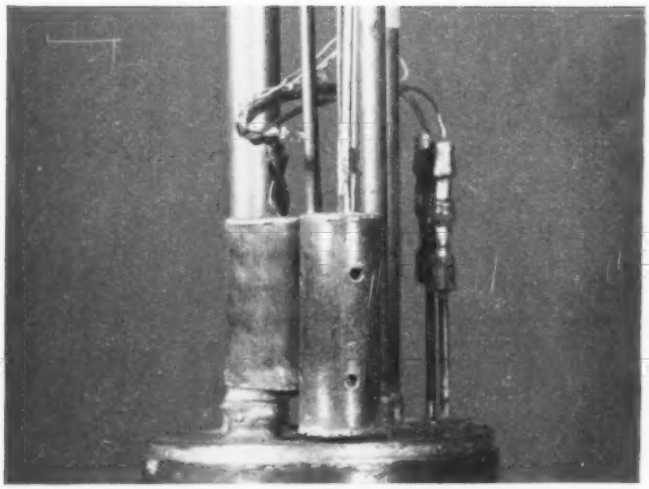


FIG. 5. Superconducting modulator in position on a cryostat. The transformer is on the left and the modulator on the right, both in superconductive lead cases to minimize pickup and radiation respectively; approximately $\frac{1}{3}$ natural size.

or reverse the e.m.f. (e.g. with a superconducting reversing switch) to identify the signal in the presence of spurious pickup, though such effects are kept to a minimum by careful design and by the use of superconductive shielding. It is usually most convenient to use a null method of measurement, in which a measured current is passed through a small series resistance to balance the unknown e.m.f. from the specimen.

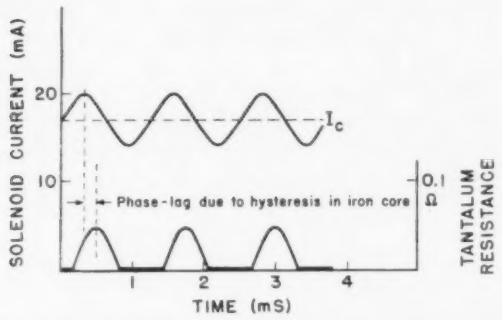


FIG. 4. Current and resistance wave form in the superconducting modulator.

Nothing has been said so far of one drawback, familiar to those working in this field. This is the occurrence of long L/R time constants in such low resistance circuits. It appears in the previous case that careful transformer design avoids any loss due to this effect, even though a small fraction of a microhenry could cause trouble at 800 c/s. However, difficulties may arise in any device which depends for its operation on direct measurement of the magnetic field produced by current in a low resistance circuit. The most recent version of a superconductive galvanometer, due to PIPPARD and PULLAN, illustrates this. In their case, they used what was effectively a tangent galvanometer, with a suspended magnet measuring the field from a pair of one-turn Helmholtz coils of lead-cored copper. This would detect a current of 10^{-5} α flowing in a circuit of resistance 10^{-7} Ω, i.e. an e.m.f. of 10^{-12} V, but since the inductance of the coils was 1·2 μH the system had a time-constant of 12 sec.

It would of course be most useful to have an actual amplifier operating in the liquid helium bath, particularly a d.c. amplifier. The high slope of the superconductive-resistive transition in a magnetic field appears to offer an ideal means of achieving such amplification. To this end, the circuit of Fig. 6 was proposed and tested. Since the transition is as susceptible to changes of *temperature* as to changes of magnetic field, it was obviously desirable to use a bridge or at least a balanced system in which the effects of temperature change would be minimized. The superconductive bridge amplifier is, in fact, a modified superconducting reversing switch. In the switch, the superconductive elements are either on or off, whereas in the amplifier each element is held, by suitable variation of the current in the control coils, approximately halfway through its transition. The resistance bridge thus formed is fed with a suitable small current and its balance is indicated by some external device, such as a galvanometer amplifier. An additional coil, of superconductive (e.g. columbium) wire is wound over each control coil, so that if these auxiliary coils are connected in the correct sense, the passage of a comparatively small current through them will unbalance the bridge. It is, of course, still necessary to reverse this current for identification.

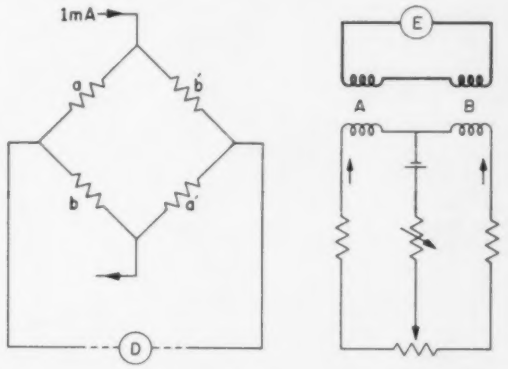


FIG. 6. Superconductive bridge amplifier. The bridge is balanced and set to a condition of maximum sensitivity by varying the control currents. The elements a, a' are controlled by coils A; elements b, b' by B.

By passing current from an external (high impedance) source through the auxiliary windings. it was found that currents as small as 10^{-7} α could be detected, which would suggest that an e.m.f. of 10^{-13} V could be detected in a circuit of resistance 10^{-6} Ω. It was soon realized, though, that

the comparatively high inductance of the circuit would render it virtually useless when connected to a low resistance source of current, since its time constant would become prohibitively long unless some appropriate negative feedback system were devised. However, the principle still applies, and work is going ahead on an evaporated film version of the device. It may be possible to use a lead–indium alloy for the controlled elements, with lead for the control wires. The first experiments are aimed at producing an evaporated version of the reversing switch.

It might be of interest here to mention another type of superconductive device. The disappearance of resistivity at the superconductive transition is accompanied by a reduction of the interaction of electrons with the lattice and hence a reduction of the electronic contribution to thermal conductivity. This becomes more marked as the temperature is lowered until, at temperatures remote from the transition, the thermal conductivity in the superconductive state is several orders lower than in the normal state. It is then possible to use the magnetic field transition to achieve thermal switching, that is, to make and break a thermal contact. This is illustrated very well in Fig. 7 (for which I am indebted to Dr. D. L. MARTIN of our laboratories) which shows the operation of a lead (Pb) thermal switch at just over 1°K in a He^3 calorimetric apparatus.

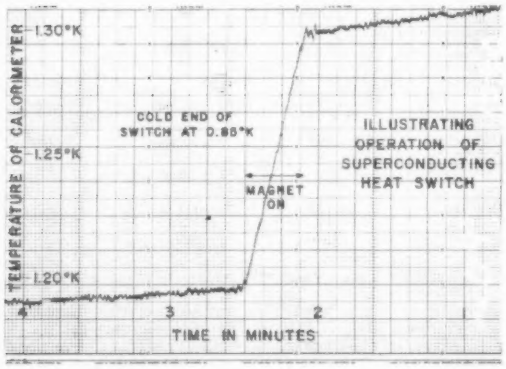


FIG. 7. Temperature–time record to illustrate operation of thermal switch.

REFERENCES

1. D. K. C. MACDONALD, W. B. PEARSON and I. M. TEMPLETON, *Phil. Mag.* **3,** 657 (1958).
2. D. K. C. MACDONALD, W. B. PEARSON and I. M. TEMPLETON, *Phil. Mag.* **3,** 917 (1958).
3. D. K. C. MACDONALD, W. B. PEARSON and I. M. TEMPLETON, *Phil. Mag.* **4,** 380 (1959).
4. D. K. C. MACDONALD, W. B. PEARSON and I. M. TEMPLETON, *Proc. Roy. Soc.* A**256,** 334 (1960).
5. R. G. CHAMBERS, *Proc. Roy. Soc.* A**238,** 344 (1956).
6. I. M. TEMPLETON, *J. Sci. Instrum.* **32,** 172 (1955).
7. I. M. TEMPLETON, *J. Sci. Instrum.* **32,** 314 (1955).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 261–272. Printed in Great Britain

PHYSICS AND CHARACTERISTICS OF THE CROSSED-FILM CRYOTRON—A REVIEW

V. L. NEWHOUSE*, J. W. BREMER† and H. H. EDWARDS‡

INTRODUCTION

THIS paper applies recently published results on the physical properties of thin superconductive films to an analysis of the Crossed-Film Cryotron (CFC). In order to avoid the danger of too much overlap with the papers presented at this conference by workers from other laboratories, the work described is mainly that done at the General Electric Research Laboratory, although work done elsewhere will be referred to.

After a brief description of the principles of the CFC, methods for studying and predicting the low-frequency properties of the CFC which depend upon joule heating will be described. Following this, the current gain of the CFC in the region where joule heating is unimportant will be analyzed. The results of the analysis are found to agree with experiment. Finally, an analog storage circuit is described which has been used to measure the time constant of one CFC driving another. Again good agreement between analysis and experiment is found.

THE CROSSED-FILM CRYOTRON[1–4]

This device is shown in Fig. 1 and typically consists of a tin "gate" film crossed by a much narrower "control grid" usually consisting of a lead film. This Sn–Pb CFC is operated just below the critical temperature of tin and far below the critical temperature of lead. Because of the temperature and the maximum current level used, the lead grid remains superconductive at all times. The tin gate will be superconducting only if the gate current is less than its "critical current", I_c. However, the tin gate can also be made resistive by the passage of a current equal to or larger than the critical control current, i_g, through the control grid.

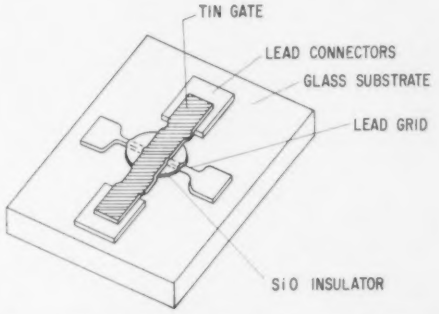


FIG. 1. Structure of crossed-film cryotron (no shield). Typical dimensions: Gate film—0·3 $\mu \times 2$ mm; insulator —0·4 μ; grid film—1 $\mu \times 25$ μ.

If I_c is larger than i_g, the CFC has gain; that is, one CFC can drive another CFC without any intermediate amplification. This condition can be achieved, as is shown below, if the CFC gate width is sufficiently larger than the grid width.

SUPERCONDUCTING–NORMAL TRANSITIONS IN TIN FILMS

Before the CFC is analyzed, some of the properties of the superconducting-to-normal transition of the tin gate films will be described.

Current-induced transition[5]

The variation of resistance of a tin film with a slowly increasing current is strongly dependent on joule heating.[6] This is illustrated by Figs. 2 and 3 which represent two extreme cases. Both cases are for tin films with a nominal thickness of 0·3 μ. In Fig. 2 we have a high-resistivity film deposited

*General Electric Research Laboratory, Schenectady, New York.

†General Electric Computer Department, Palo Alto, California.

‡General Electric Light Military Electronics Department, Ithaca, New York.

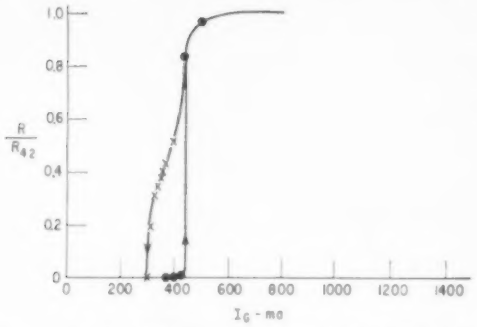


FIG. 2. Current-induced transition for tin film on glass substrate. (ρ = 1·28 $\mu\Omega$-cm, T = 3·74°K, T_c* = 3·87°K) ● Increasing current; × decreasing current.

on glass which is a comparatively bad conductor of heat, while Fig. 3 represents a low-resistivity film, obtained by high-speed deposition, on a single-crystal sapphire which is a very good conductor

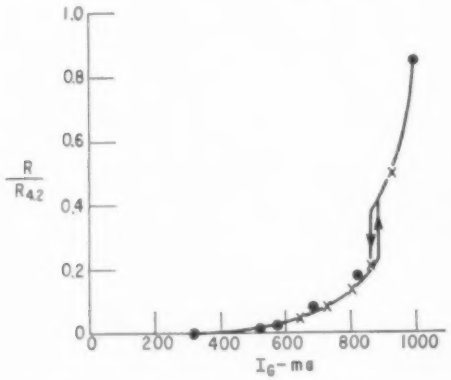


FIG. 3. Current-induced transition for low-resistivity film on sapphire substrate. (ρ = 0·30 $\mu\Omega$-cm, T = 3·77°K, T_c = 3·84°K) ● Increasing current; × decreasing current.

of heat. An intermediate case, a high-resistivity film on sapphire, is shown in Fig. 4. All three figures show definite hysteresis loops with sharp transitions from a low- to a high-resistivity state.

* In this paper T_c is defined as the temperature where the superconductor has fallen to one half of its normal resistance in the absence of a magnetic field. For the films used, this differs by about 0·01°K from that temperature at which zero resistance is obtained. Refs. 1, 5 and 12, by two of the authors, have used the latter definition for many of their figures.

These transitions are caused by thermal heating, not by the superconducting properties of the film. The sharpness of the transition is caused by a sudden rise in temperature of the film as explained below.

Heating effects can be minimized by applying

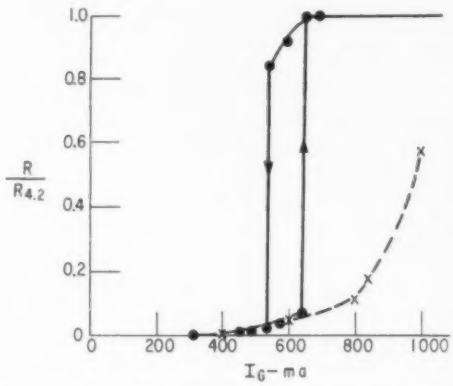


FIG. 4. D.C. and pulse current-induced transitions for tin film on sapphire substrate. (ρ = 1·1 $\mu\Omega$-cm, T = 3·77°K, T_c = 3·86°K) ● Direct current; × 4 μsec pulses.

the current in pulses.[5] In this case smooth curves such as those shown in Fig. 5 are obtained. Joule heating effects are minimized and perhaps completely absent. These curves will therefore be referred to as "isothermal" transitions, where the transition is mainly due to the electromagnetic effect of the current, thermal properties of the substrate and helium, and specimen geometry. From such curves, it is possible to derive the d.c. hysteresis loops on a sapphire substrate.

Derivation of d.c. hysteresis loop

Consider a current I flowing in a tin film which has been deposited on a thin sapphire disk whose area A is only a little larger than that of the tin film. The film has a resistance R which is a function of I and of the film temperature T. T is determined by joule heating (I^2R) and the heat loss to the liquid helium. Heat flow from the surface of the sapphire substrate to the liquid helium is much less efficient than heat flow through the sapphire. Because of this and the relatively small area of the substrate, it is justified to assume that the

sapphire will be at the same temperature T as the film. The helium temperature is T_0. The heat loss to the helium will take place from the film of area A directly to the liquid, and through the substrate to the liquid. Hence the total heat-loss area will be $2A$. The heat loss to the helium will be $2FA(T-T_0)$ where F is a constant. For a steady current I, the film resistance and temperature will adjust themselves until thermal equilibrium is obtained. This condition is given by

$$I^2R(T, I) = 2FA(T-T_0) \qquad (1)$$

It is assumed that the resistance of the film does not vary with position. This is true provided the temperature difference between the middle and the ends of the film is much less than the width (approximately 0·02°K) of the R vs. T transition for constant current. It can be shown that this condition is fulfilled in the case described below.

Equation (1) can best be solved graphically. Isothermal transition curves of R vs. T are first obtained experimentally, such as those shown in Fig. 5. These are then plotted as I^2R vs. T as shown schematically in Fig. 6. The equilibrium condition of equation (1) is then given by the intercepts of the straight line of slope $2FA$ with the curves, also shown in Fig. 6.

When the current is increased from i_0 the thermal-equilibrium position is given by the lowest intersection of the corresponding current curves with the straight line. For a current just above i_3 an unstable equilibrium is reached, and

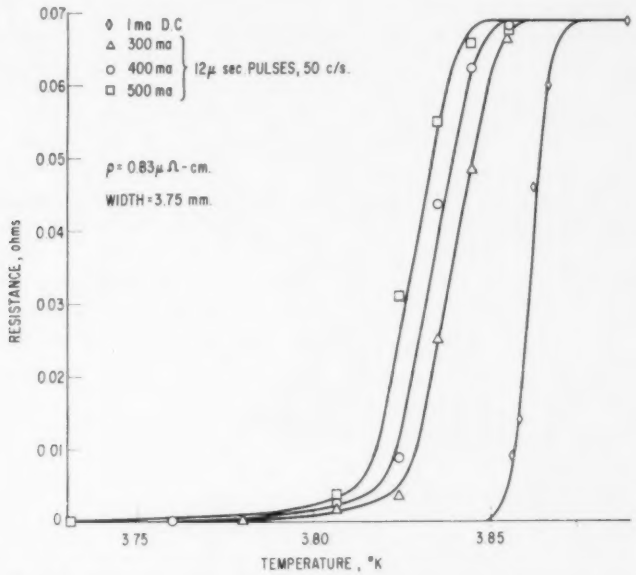


FIG. 5. Film resistance as a function of current and temperature in the absence of joule heating.

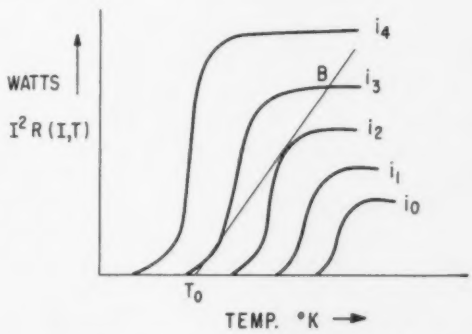


FIG. 6. Joule heat generated in film as a function of film temperature and current I. Straight line represents $FA(T-T_0)$, the heat loss to bath.

the resistance and temperature change in a discontinuous fashion to the values given by the upper intersection B. If the current is now decreased below i_2 the equilibrium again becomes unstable and returns discontinuously to a point close to T_0. Clearly snap transitions occur when the straight line is tangential to the I^2R curve, i.e. when

$$I^2\left(\frac{\partial R}{\partial T}\right)_I = 2FA \qquad (2)$$

It can also be seen that for current increasing from zero, the resistance change due to Joule heating will remain negligible as long as

$$I^2\left(\frac{\partial R}{\partial T}\right)_I \ll 2FA$$

To compare equations (1) and (2) with experiment, a d.c. transition curve was constructed from the data of Fig. 6. This is compared with a curve obtained by direct measurements in Fig. 7.

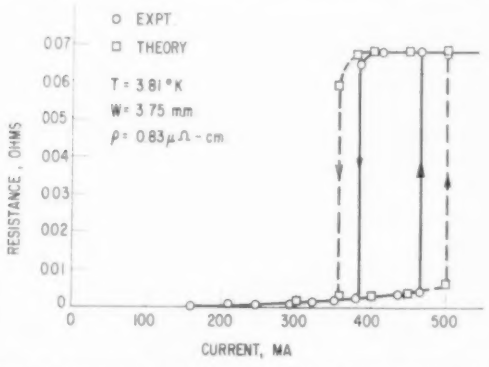


FIG. 7. Comparison of d.c. transition curve, constructed from isothermal data of Fig. 5, with experiment. $\Delta T = 0{\cdot}065°K$.

The agreement is seen to be reasonable. The value of F used was $0{\cdot}2$ W/cm^2 °K as given by KARAGOUMIS[7].

We can conclude therefore that, at least in the case of a film on a high thermal conductivity substrate, we can calculate the d.c. R vs. I hysteresis loop from the isothermal curves. The above treatment is not justified in the case of a substrate with a low thermal conductivity such as glass because the temperature is not then sufficiently uniform throughout the film. It is found, in fact, that the shape of the R vs. I hysteresis loop for tin films on glass is strongly dependent on the ratio of film width to length. This is not the case for tin films on sapphire.

The behaviour of tin and indium films under pulse conditions over a wide range of temperature and film thickness has been studied by CRITTENDEN and his co-workers[8,9]. Thermal relaxation times of lead and tin films have been studied by them and by BROOM and RHODERICK[10] and BROOM and SIMPSON[11].

Thermal nucleation and propagation[12]

Figs. 2–4 and 7 illustrate that a tin film carrying a current can have two equilibrium values of temperature and resistance, depending upon its history. It is interesting to observe how the thermal transitions between these two states occur. It has been shown that a transition from a low- to a high-resistance state can be initiated by a nucleation process. The situation is described in Fig. 8 which represents a current-carrying wire

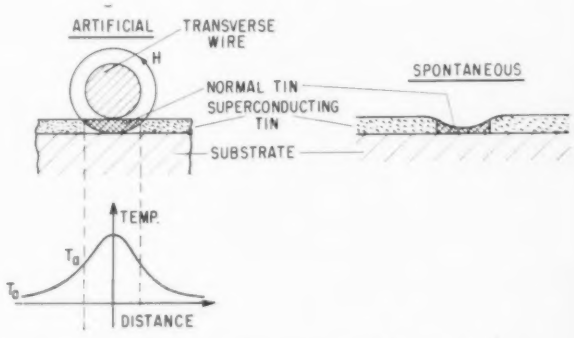


FIG. 8. Artificial and spontaneous models for nucleation of resistive regions.

above an insulated film. The magnetic field from the wire produces a resistive region in the film underneath it. If a current is now passed through the film, joule heat will be developed in the resistive region and the temperature distribution will be somewhat as shown. If the current through the film becomes high enough the rise in temperature will increase the resistive region and this in turn will increase the heat developed. Under suitable conditions the resistive region should grow spontaneously at a speed determined by the heat

developed, I^2R, and by the thermal properties of the film and substrate.

These ideas are confirmed by experiment as shown in Fig. 9. The R vs. I loop is for a tin film on a glass substrate. It is shown that a transition can be initiated by passing a current I_N through an insulated lead wire crossing the film which is carrying a current I_M.

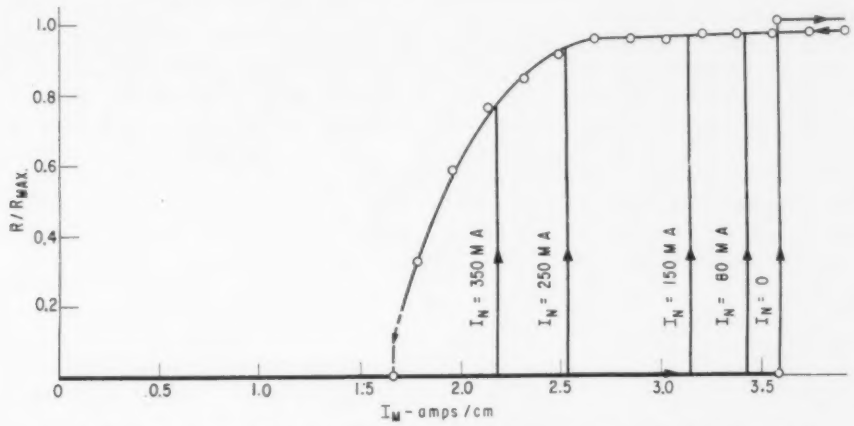


FIG. 9. Nucleation current for a tin film on a glass substrate, $\Delta T = 0{\cdot}29°K$, film width = 0·28 mm, wire diameter = 75 μ.

The dynamics of this process have been studied with a specimen crossed by two "grid" wires as shown in Fig. 10. A broad pulse is passed through the tin film and a much narrower nucleation pulse through one or both of the transverse wires. The voltage variation with time is shown in Fig. 11(a) for the case where wire 1 (Fig. 10) is used for nucleation. If we assume that the normal–superconductive boundaries are at all times parallel to the wire, and move at constant speed, then we would expect a break in the voltage–time curve when one of the boundaries reaches the nearer end of the film. If the nucleating wire is $1/4\ l$ from the end of a film of length l, then some consideration will show that the break in the V–T curve should occur when the voltage has reached half its maximum value. Fig. 11(a) shows that this is just what happens. Fig. 11(b) refers to the case when the central nucleating wire is used. Here the break in the V–T curve occurs just before V reaches its maximum value. This indicates that either the wire was not exactly centered, or that the speed of the two normal–superconductive boundaries was not quite equal. Fig. 11(c) refers to the case where both nucleating wires were pulsed simultaneously. For further details reference should be made to the original paper.[12]

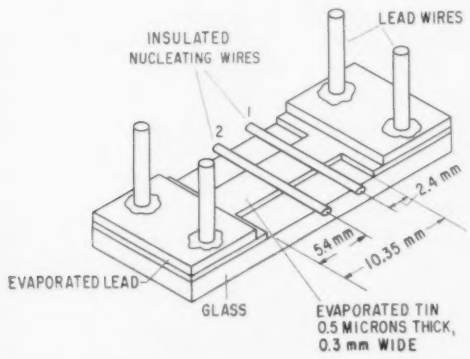


FIG. 10. Specimen structure. Substrate thickness = 0·25 mm. Lead electrode thickness = 10 μ. Diameter of nucleating wires = 75 μ. Tin film resistivities at $4{\cdot}2°K = 1{\cdot}1 \pm 0{\cdot}1\ \mu\Omega$-cm. Tin film critical temperatures = $3{\cdot}85 \pm 0{\cdot}05°K$. The lead wires and electrodes remain superconducting throughout, thus avoiding contact heating.

If the transverse wires are not pulsed, thermal propagation will start spontaneously when the current reaches a high enough value. By observing the voltage across the film as a function of time, we have found that the propagation process

normally starts in the center of the film where the temperature is highest. However, if the film has a narrow region where the current density is high, then the propagation will start at that point. The distance of the nucleation site from the nearest end of the film can be deduced from the fraction of the maximum voltage at which the break in the V–T curve occurs.

The above results show that thermal propagation acts as a means of amplifying a brief current impulse, and as a means of locating in which of many transverse wires the impulse occurs. It can thus be used in output elements, digital analog converters and in any application where the long thermal time constant is acceptable. It has been shown that the propagation speed depends on the thermal conductivity of the substrate. Speeds of 1000 cm/sec have been reported.[9]

Film critical current

Reference to Fig. 4 shows that the current I_c at which resistance first appears is much smaller than the current at which the resistance reaches its maximum value. This is not surprising, because

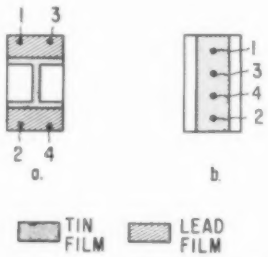


FIG. 12. Sample geometry. Current leads—1,2; voltage leads—3,4. (a) Sample shape for current transition measurements. (b) Sample shape for magnetic-field transition measurements.

in a superconducting thin film, the surface current density is much higher at the edges than in the middle. Thus it is to be expected that a superconducting-to-normal (SN) transition will start at the edges of the film for very small currents. When the film SN transition approaches completion, the current distribution must approach uniformity. Hence for films much thicker than the penetration depth, the SN transition should be complete when the surface field of an assumed uniformly distributed current approaches the bulk critical field H_c.

For flat films, much thinner than the penetration depth, GINZBURG[13] has derived the expression

$$\frac{0{\cdot}2\pi I}{W}H_c = \frac{4}{3}H_B^2 \qquad (3)$$

for the critical current, neglecting the effect of the edges. Here H_B is the bulk critical field, I the critical current in a film of width W (assumed uniformly distributed) and H_c the critical field of the film, measured with a field parallel to the surface of the film. It is clear that as maximum film resistance is approached, the current distribution approaches uniformity, and in that case equation (3) should be applicable. We have in fact been able to show[5] that equation (3) is obeyed when I is taken as the pulse current which gives half the maximum resistance, and H_c as the field which gives half the maximum resistance.

It should be pointed out that a unique value of H_c cannot be defined for most flat films because of a thinner region near the film edges which has a higher critical field than the thicker region in the center. A specimen geometry has, therefore, to be chosen, such as shown in Fig. 12(b), which prevents the film edges from short circuiting the voltage measurement points.

The magnitude of critical current I_c at which resistance first appears cannot as yet be calculated. However, experimentally in the case of $0{\cdot}3\,\mu$ films of tin, it has been shown to vary linearly with film widths from between 0·01 cm to 0·4 cm. This suggests that any structure of the intermediate state in films under these conditions must be small compared to 0·01 cm. It is shown below that in the case of a film deposited on top of an insulated superconducting "shield" plane, the current density near the edges is reduced. We would expect, therefore, that I_c should increase and this is found to be the case. It is also found that, for shielded $0{\cdot}3\,\mu$ tin films, $I_c \propto (T_c - T)^2$ for values of T close to T_c.

We can summarize the results of this section as follows. From an engineering point of view the value of current I_c at which resistance first appears is of paramount interest. Experiment shows that $I_c \propto W$, and that $I_c \propto (T_c - T)^2$. Theory is not at present able to set a lower limit to I_c but it can

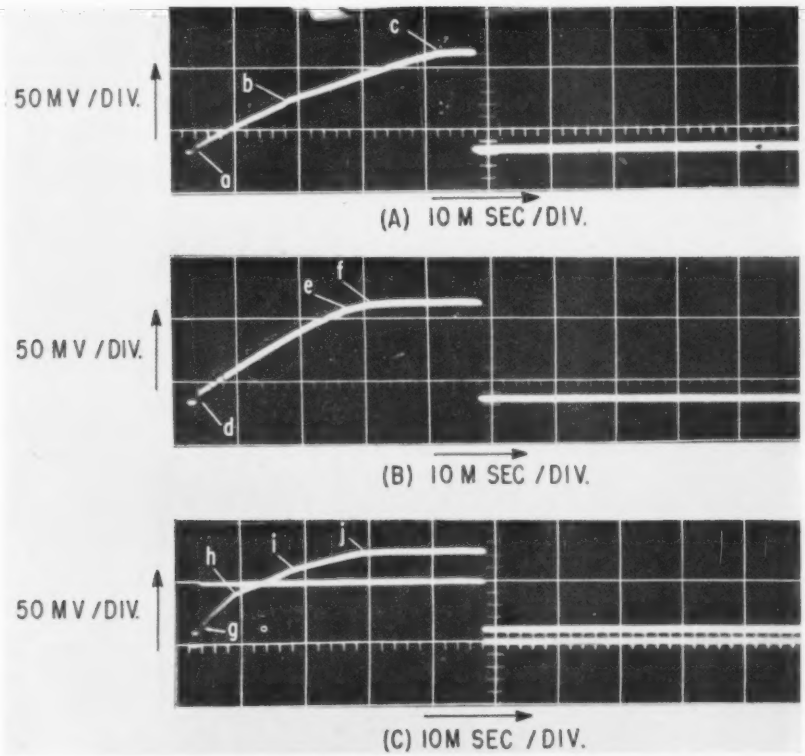


FIG. 11. The voltage across a film as in Fig. 10 associated with the growth of artificially nucleated resistive regions within it. (A) Nucleus initiated under wire 1. (B) Nucleus initiated under wire 2. (C) Nucleus initiated under wires 1 and 2 simultaneously (this is superimposed on the voltage across a $\frac{1}{2}$ Ω resistor carrying the maintaining current). The maintaining current pulse was 100 mA for 45 msec. The nucleating current pulse was 775 mA for 0·76 msec. The bath temperature was 3·57°K.

[*facing p.* 266

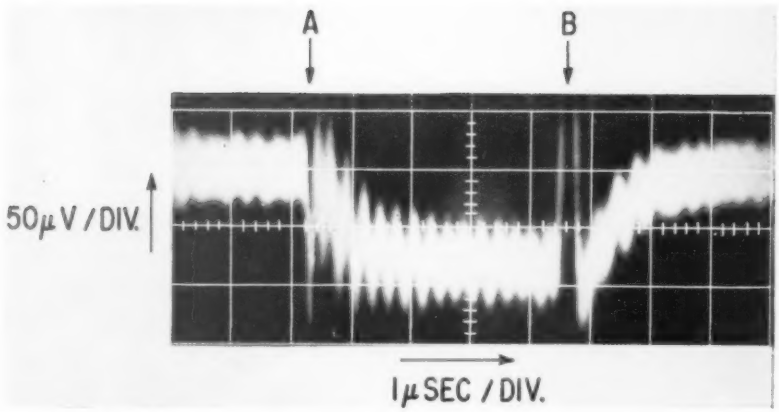


FIG. 17. Output voltage of cryotron. I_0 and I_{out} (See Fig. 16) are on at the start of the sweep. I_{in} is switched on at A and I_0 is switched off at B.

set an upper limit, given by equation (3), and this can be used in determining an upper bound to cryotron performance as will be shown below.

CFC ANALYSIS[14]

With the above facts about the superconducting-to-normal transitions in films understood, we can now return to the analysis of the CFC.

Magnetic field—analysis of the CFC structure

To approach this problem we first consider the case of an infinitely thin current-carrying superconducting strip a distance d above a superconducting shield plane. It is well known[15,16] that screening currents will be induced in the shield plane which will increase the field between the film and the plane and decrease it everywhere else. The field distribution near a strip of width w has been analyzed using potential theory, and the results are shown graphically in Fig. 13.*

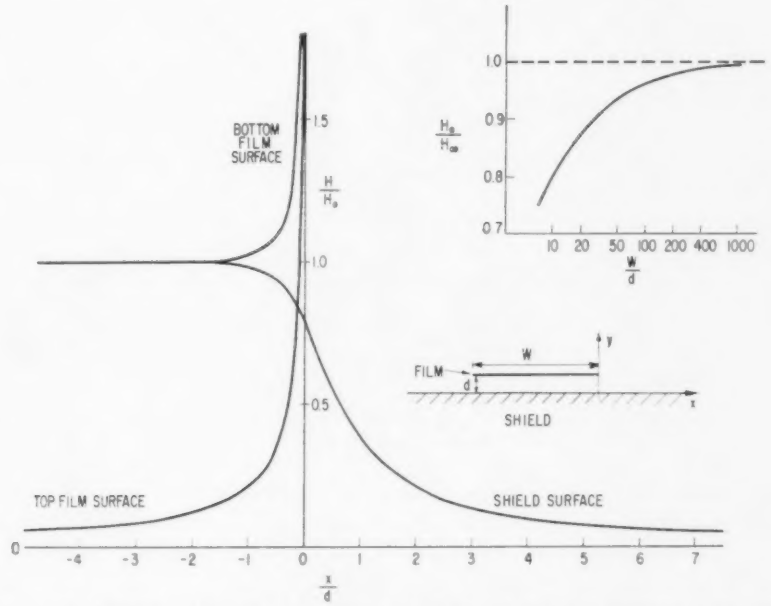


FIG. 13. Surface fields for a superconducting film of width w carrying a current I, over an infinite superconducting shield plane. ($H_\infty = 0{\cdot}4\,\pi I/w$.)

In the figure, H_0 is defined as the field in the midpoint of the region between the film and the shield. We see that for $w \gg d$ the field between the film and the shield is very nearly constant and approximately equal to H_0. The value of H_0 can be derived very simply by considering the limit where $d/w \to 0$. In that case the field above the film is zero. By integrating around the film and using the formula $\oint H \cdot ds = 0{\cdot}4\,\pi I$ we obtain

$$(w \times H_0) + (w \times 0) = 0{\cdot}4\pi I$$

i.e.

$$H_0 = 0{\cdot}4\pi I/w \qquad (4)$$

Consider now the case of an unshielded CFC. If the grid is carrying a current which is insufficient to make the gate resistive, then the field configuration around the gate film will be the same as that around the film in Fig. 13, since the gate acts as the shielding plane. In particular the maximum field applied to the gate will be H_0

* It is assumed that $d \ll w$ and that $t \ll d$ where t is the film thickness. However, as long as the first assumption is correct, the results are not strongly affected even if t is of the same order of magnitude as d. Also it is assumed that the film penetration depth and the shield penetration depth are both negligible compared to d.

and will occur below the midpoint of the grid. Hence the grid current i_g at which resistance will first appear in the gate will be, from (4), such that

$$H_c = 0{\cdot}4\pi i_g/w \qquad (5)$$

where w is the grid width.

In the case of a shielded CFC the same situation holds. Hence, i_g should be the same for a shielded or an unshielded CFC and should be given by equation (5). This is confirmed by curve 3 in Fig. 14 where H_c, calculated from measured values of i_g, is compared for various shielded and unshielded CFC's at various temperatures. The field for unshielded cryotrons is calculated only for the initial resistance curve, because of the difficulties associated with magnetic-field calculations once the gate has gone partially resistive.

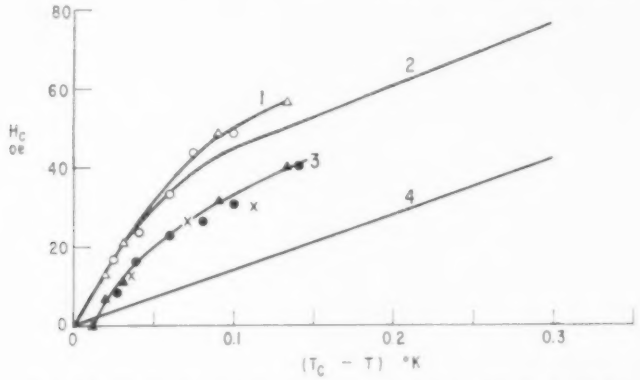


FIG. 14. H_c for tin as a function of temperature. (1) H_c for 0·3 μ tin gate film calculated from grid-current measurements at 1/2 resistance, △ 80 μ grid width, ○ 25 μ grid width. (2) H_c for 0·3 μ tin film measured directly [geometry of Fig. (12b)]. (3) H_c calculated from grid current just sufficient to produce gate resistance, using equation (5). ▲ 80 μ grid width (shielded gate), ● 25 μ grid width (shielded gate) × 16 μ grid width (unshielded gate). (4) H_c for bulk tin. (The intercept of curve 3 with the abscissa corresponds to the width of the film-resistance transition with temperature at zero field.)

CFC characteristics

D.C. characteristics for typical shielded and unshielded CFC's are compared in Fig. 15. The solid lines are for a shielded CFC and the dashed lines are for an unshielded CFC. The left-hand ordinate refers to the two curves which intersect it (i.e. at 91 mA and 340 mA gate current). The right-hand ordinate is for the other four curves. Three points of interest are:

(1) i_g is the same for the shielded and the unshielded unit. This was explained above.

(2) The resistance of the shielded CFC is constant at high grid currents. This is clearly due to the fact that, as shown in Fig. 13, the field falls rapidly with distance from a shielded grid. Hence the maximum gate resistance cannot greatly exceed that corresponding to a length w of gate film, where w is the grid width. The maximum resistance calculated in this way is shown in the figure and is seen to agree well with experiment.

(3) The critical current of the tin gate is much larger for the shielded than the unshielded case. This can be understood qualitatively because the superconducting shield makes the current distribution more uniform. However, a quantitative understanding is lacking.*

D.C. gain of the CFC

Since the critical current of a tin gate varies directly with the width we can define a quantity j_c

* The explanation advanced in our Ref. 12 is incorrect. It was based upon a misunderstanding of the terminology of GINZBURG's[13] paper.

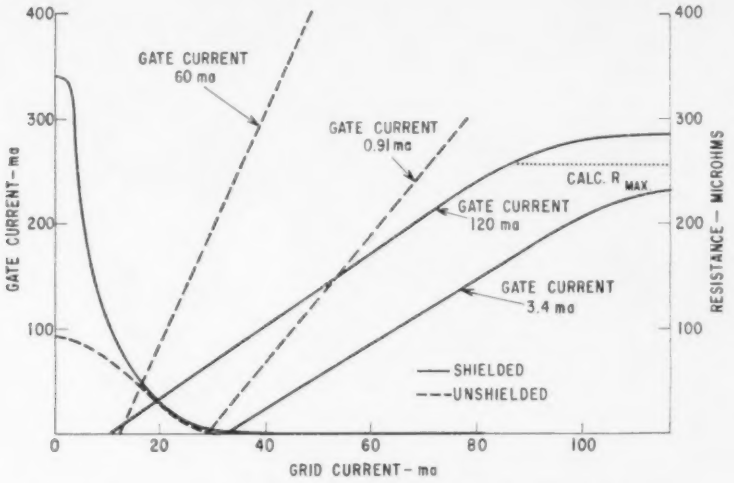


FIG. 15. Comparison of electrical characteristics of unshielded and shielded CFC. Shield insulation thickness = 0·4 μ. Grid width—30 μ; gate width—2 mm; $T_c - T = 0{\cdot}08°K$.

equal to the critical current per unit width of gate film. Hence for a gate of width W,

$$I_c = j_c W \tag{6}$$

The gain will be defined as

$$g = \frac{I_c}{i_g} \tag{7}$$

where i_g as defined above is the minimum grid current at which the gate becomes resistive for low gate currents.

Substituting from (5) and (6) into (7), we find

$$g = \frac{j_c 0{\cdot}4\pi}{H_c}\frac{W}{w} = \frac{H_i}{H_c}\frac{W}{w} \tag{8}$$

where $H_i = 0{\cdot}4\pi j_c$ is the surface field between a gate film and shield when the gate film is carrying a current j_c per unit width.

Measurement of time constant[16]

If current is pulsed through a CFC grid, the gate resistance appears in a time of the order of milli-microseconds.[4] Since this is much faster than the electrical time constant, the speed limit of circuit operation is set by the cryotron gate resistance and the inductance of the superconducting load. A convenient circuit for measuring time constants is shown in Fig. 16(a) with the equivalent circuit in Fig. 16(b). Its operation is as follows.

If a current I_0 is injected between X and Z, most of it will flow through XZ rather than XYZ because $L_2 \gg L_1$. If a current is now passed through the input grid, I_0 will start to become diverted from XZ to YYZ with a time constant $(L_1+L_2)/R$. (R is the resistance produced in XZ by the input grid current.) The current through XYZ can be monitored by passing d.c. through MN and observing the resistance. The rise in voltage across MN for such a circuit is shown in Fig. 17.* Eventually all of I_0 will be diverted through XYZ. If I_0 is now switched off, a circulating current will be produced of magnitude $I_{circ} = I_0 L_2/(L_1+L_2)$. We have "stored" such currents for several hours. If the input grid current is still on when I_0 is swtiched off, then I_{circ} will decay with the same time constant $(L_1+L_2)/R$. This is illustrated in Fig. 17. Owing to the low output voltage across MN, amplifier noise makes

* The actual circuit used was a loop of the CFC shift register described by EDWARDS *et al.*[17]

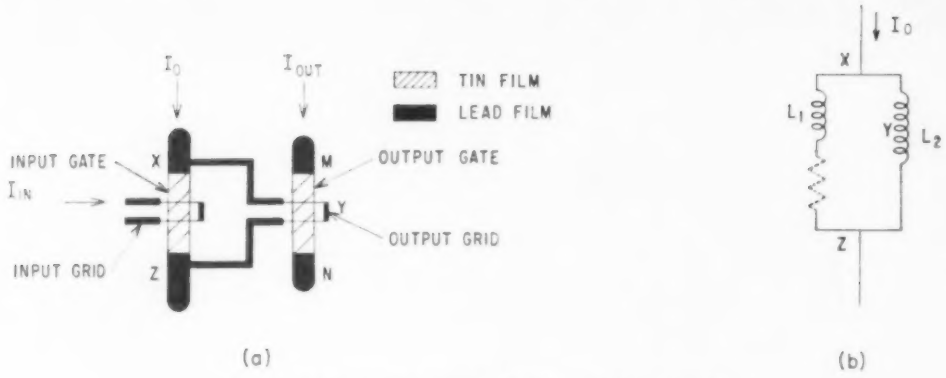


FIG. 16. (a) Cryotron storage cell; (b) equivalent circuit.

it difficult to measure accurately the time constant from pictures such as Fig. 17. We have therefore used a method which depends on the fact that as soon as the current through the input grid is turned off, the currents through XZ and YXZ will remain "frozen". [This is due to the fact that as $R \to 0$, the time constant $(L_1+L_2)/R \to \infty$.] The result of pulsing the input grid in the presence of a stored current is shown in the decaying curve of Fig. 18. Under these conditions, the voltage after each pulse can be read accurately with a d.c. microvoltmeter. Current build-up can also be observed, as shown in Fig. 18, by pulsing the input grid in the presence of a steady current I_0. Both these curves can be fitted to exponential curves with a time constant of 0·38 μsec. The value calculated from the measured resistance and the calculated circuit inductance is 0·33 μsec. Fig. 17 gives approximately the same value. We can deduce therefore that heating effects do not play an appreciable part in the operation of this circuit, and the measured d.c. CFC characteristics are the same as the a.c. characteristics. Furthermore, the circuit of Fig. 16 is clearly well suited for analog as well as digital storage.

CFC gain-bandwidth product

The time constant of the circuit of Fig. 16 is a convenient figure of merit, because one CFC drives one other CFC. The time constant is

$$\tau = (L_1+L_2)/R \tag{9}$$

The inductance of a film of width w separated by a distance d from a superconducting shield plane can be shown[15] to be $L = 4\pi\,(d/w)\times 10^{-9}\,H$. Clearly since $w \ll W$ (where W is the CFC gate width), $L_1 \ll L_2$. From (9), $\tau \sim L_2/R$

$$L_2 = 4\pi\frac{dW}{w}\times 10^{-9}H \tag{10}$$

The resistance of that portion of the gate film of resistivity ρ, thickness t and width W under a grid of width w is

$$R = \frac{w\rho}{tW} \tag{11}$$

Hence from (10) and (11)

$$\tau = L_2/R$$
$$= 4\pi\frac{dt}{\rho}\left(\frac{W}{w}\right)^2\times 10^{-9}\ \text{sec} \tag{12}$$

This assumes that the grid current is large enough to produce the maximum gate resistance. Let us assume that the grid current required for this is $K\,i_g$ where i_g is the minimum grid current required to produce gate resistance as defined by equation (5). We can now define a new gain G, as the ratio of the gate critical current to the minimum grid current giving full resistance. From (8) we have

$$G = \frac{1}{K}\frac{H_t}{H_c}\frac{W}{w} \tag{13}$$

Substituting for W/w from equations (13) and (12)

we finally obtain

$$\tau = 4\pi \frac{dt}{\rho}\left(\frac{H_c}{H_t}\right)^2 (GK)^2 \times 10^{-9} \text{ sec} \qquad (14)$$

There are four points of interest in (14).

(1) $\tau \propto t(H_c/H_t)$. This shows that there is an optimum value of the film thickness t, because as t increases, (H_c/H_t) decreases towards unity. For the $0{\cdot}3\,\mu$ films presently used (H_c/H_t) = between 15 and 50, hence there appears to be plenty of opportunity for reduction of τ.

(2) (H_c/H_t) decreases with temperature. Hence τ must do the same.

(3) Since G is set by circuit considerations, and assuming that K is not a function of gate and grid widths, neither is τ. Hence τ cannot be made smaller by reducing the CFC area but only by reducing the grid–shield spacing.

(4) The CFC bandwidth is proportional to G^{-2}.

We can compare equation (14) with the experimentally obtained values of time constants of approximately $0{\cdot}4\,\mu$sec. Present values of the material constants are: $t = 0{\cdot}3\,\mu$, $d = 0{\cdot}4$–$1{\cdot}0\,\mu$, $\rho = 3$–12×10^{-7} Ω-cm, $H_c/H_t = 15$–50, $K = 2$–5. A practical value for $G = 2$.

Substituting these values into (14), we obtain a present lower limit for τ of the order of $0{\cdot}1\,\mu$sec. Using equation (3) it is possible to show that for a film with an ideally sharp transition, $H_c/H_t \sim 3$ at the presently used operating temperature of $T_c - T \sim 0{\cdot}08$°K. Under these conditions equation (14) could lead to a time constant of under 3 mμsec. Other, and perhaps easier, means of reducing the time constant are to use lower temperatures and thinner insulation, and to obtain higher gate film resistivities.

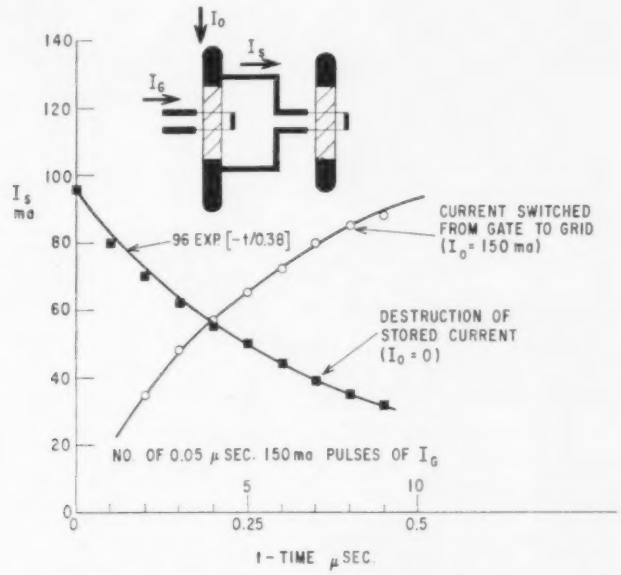


FIG. 18. Change of sorted current due to $0{\cdot}05\,\mu$sec pulses on input grid of storage cell. Curves are fitted to the data points corresponding to a time constant of $0{\cdot}38\,\mu$sec.

SUMMARY AND CONCLUSIONS

This review has discussed thermal nucleation and propagation phenomena in superconducting films which should be useful in various input–output and analog–digital converter applications. It has been shown that cross-film cryotron static and dynamic characteristics can be analyzed

using standard electromagnetic theory and that good agreement with experiment is obtained. However, the factors which control the width of the current-induced transition in the absence of joule heating, remain to be investigated. Increased understanding of these phenomena may make it possible to reduce present values of H_c/H_i thus increasing CFC bandwidth. The time constant for the CFC is independent of its area if the gain and temperature are held constant. With present materials τ lies between 0·1 and 0·4 μsec. for $T_c - T \sim 0{\cdot}08$°K.

REFERENCES

1. V. L. NEWHOUSE and J. W. BREMER, *J. Appl. Phys.* **30,** 1458 (1959).
2. V. L. NEWHOUSE and J. W. BREMER, *Electron Device Research Conference, Ithaca, N.Y., June,* 1959.
3. M. L. COHEN and co-workers, *Electron Device Research Conference, Ithaca, N.Y., June,* 1959.
4. A. E. BRENNEMAN and R. DE LANO, *Electron Device Research Conference, Ithaca, N.Y. June,* 1959.
5. J. W. BREMER and V. L. NEWHOUSE, *Phys. Rev.* **116,** 309 (1959).
6. W. BUCKEL and R. HILSCH, *The Intermediate State of Thin Metal Films* p. 323. Proceedings of the Fifth International Conference on Low Temperature Physics and Chemistry (1957).
7. A. KARAGOUMIS, *Bull. Inst. Int. Froid Annexe Suppl.* (*Louvain Conference*) 195 (1956).
8. E. C. CRITTENDEN, JR., J. N. COOPER, F. W. SCHMIDLIN and A. J. LEARN, Critical currents in superconducting thin films, *Structure and Properties of Thin Films* (Edited by NEUGEBAUER *et al.*) p. 282. Wiley, New York (1959).
9. E. C. CRITTENDEN, J. N. COOPER and F. W. SCHMIDLIN, *The "Persistor"—A Superconducting Memory Element*. Space Technology Laboratories, Inc. Report. Also published in *Proc. I.R.E.* **48,** 1233 (1960).
10. R. F. BROOM and E. H. RHODERICK, *Phys. Rev.* **116,** 344 (1959).
11. R. F. BROOM and O. SIMPSON, *Brit. J. Appl. Phys.* **11,** 78 (1960).
12. J. W. BREMER and V. L. NEWHOUSE, *Phys. Rev. Letters* **1,** 282 (1958).
13. V. L. GINZBURG, *Dokl. Akad. Nauk S.S.S.R.* **3,** 102 (1959).
14. V. L. NEWHOUSE, J. W. BREMER and H. H. EDWARDS, *Proc. I.R.E.* **48,** 1395 (1960).
15. D. R. YOUNG, Superconducting circuits, *Progress in Cryogenics* Vol. 1, p.1. Academic Press, New York (1959).
16. V. L. NEWHOUSE, J. W. BREMER and H. H. EDWARDS, *The Crossed Film Cryotron and Its Application to Digital Computers.* Proceedings of the Eastern Joint Computer Conference, p.255 (1959).
17. H. H. EDWARDS, V. L. NEWHOUSE and J. W. BREMER, *Proceedings of the Symposium on Superconductive Techniques for Computing Systems, Washington,* 1960. Office of Naval Research (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 273–278. Printed in Great Britain

CLOSED-CYCLE HELIUM REFRIGERATION

HOWARD O. McMAHON and WILLIAM E. GIFFORD

Arthur D. Little, Inc., Cambridge, Massachusetts

INTRODUCTION

THE practicability and range of applications of superconductive circuits appear likely eventually to be determined by the ease and reliability of maintaining the required low temperature. Present knowledge indicates that a temperature near the boiling point of liquid helium, or perhaps a little lower, will be required. Since there is no known means for maintaining such a temperature other than by compression and expansion of helium, the problem is reduced to the question of the feasibility of constructing a mechanical closed-cycle helium refrigerator of sufficient reliability.

In order to liquefy helium it is necessary to compress the gas to approximately 200–300 lb/in^2, cool it to approximately 15°K or colder, and expand it through a Joule–Thomson cycle to one atmosphere or less. All liquefiers or refrigerators must necessarily have this feature in common. Refrigerators differ from one another according to the way in which they accomplish the initial cooling from 300 to 15°K. The classical method, and the one first used to liquefy helium,[1] uses liquid ammonia, sulphur dioxide or Freon, liquid nitrogen and liquid hydrogen to accomplish the necessary precooling. This of course leads to a complex, bulky machine having low efficiency, poor reliability and with very poor possibility of development in the direction of simplicity and self-sufficiency. A considerably improved method of providing precooling was developed by COLLINS and has led to the standard ADL-Collins laboratory liquefier[2,3]. In this system, portions of compressed helium gas are expanded in expansion engines at intermediate temperatures to produce the refrigeration required for cooling the Joule–Thomson stream. It is possible that the COLLINS system could provide a basis for the development of a refrigerator having the required high reliability, but it appears that there may be formidable problems connected with the refinement of the expansion engines and valve mechanisms. Moreover, the COLLINS system carries the entire helium flow through countercurrent heat exchangers, which are necessarily more bulky and costly than reciprocating-flow thermal regenerators.

A new engine-expansion system has been developed[4,5] which utilizes reciprocating flow through a thermal regenerator instead of counterflow heat exchangers. This feature permits the engine valves to be operated at room temperatures, thereby eliminating many of the mechanical problems associated with them. A distinctive feature of the new expansion engine is that enthalpy is ejected from the system in the form of heat rather than as mechanical work. This eliminates the necessity of any large mechanical forces and allows for simplicity of design and high reliability.

THE SINGLE-STAGE CYCLE

A single-stage refrigerator utilizing the new cycle is shown schematically in Fig. 1. It consists of a thin-wall stainless-steel cylinder into which is

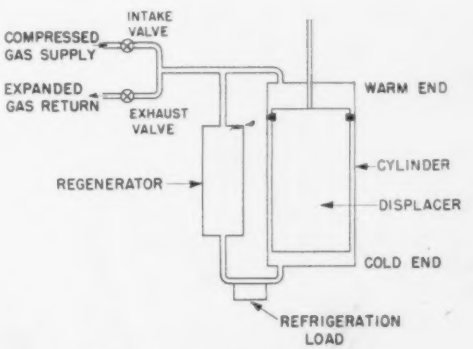


FIG. 1. Single-stage refrigerator.

fitted a displacer in the form of a fairly closely fitting plastic piston which is free to move longitudinally throughout a limited distance. A rod is attached to the room-temperature end of the displacer and is used to control its motion. The other end of the displacer–cylinder combination is at the low temperature. As the displacer is moved, the warm and cold volumes at the two ends of the cylinder are varied, one increasing at the expense of the other. However, these two volumes are connected together through a thermal regenerator having relatively low resistance to gas flow. Consequently, the pressure above and below the displacer is always approximately equal; and no appreciable force is required to move it.

The thermal regenerator consists of a thin-wall stainless-steel tube filled with circles of fine metal screen which are stacked to form a cylindrical structure just equal to the internal dimensions of the regenerator shell. The upper end of the regenerator is at approximately room temperature and the other end is at the low temperature. In spite of the large temperature gradient along the length of the regenerator, there is very little heat flow because of the poor thermal conductivity of the stacked structure.

The intake and exhaust valves, which are located at room temperature, are operated by cams attached to a crank shaft which controls the motion of the displacer. Compressed gas is supplied to the refrigerator from a more or less conventional closed-cycle compression system consisting of a reciprocating two-stage compressor and the usual ballast chambers and coolers.

The cycle may be considered to be made up of four operations: *Pressurization*, *Intake*, *Expansion* and *Exhaust*. At the beginning of the pressurization the displacer is at the extreme cold end of the cylinder so that the cold volume is a minimum and the warm volume is a maximum. The intake valve is opened, and compressed gas rushes into the warm volume (and also the free volume of the regenerator) until the pressure is equal to that of the supply. The sudden increase of pressure in the warm volume, of course, causes adiabatic heating and results in a considerable temperature rise of the gas.

With the intake valve remaining open, the displacer is moved so as to enlarge the cold volume and diminish the warm volume. This displaces the heated gas through the regenerator to the cold end causing it to decrease in volume by an amount proportional to the ratio of the absolute temperatures at the two ends. Since the intake valve is open during the operation, the pressure is constant; and more gas flows in through the intake valve and mixes with the displaced stream entering the warm end of the regenerator. At the end of the intake stroke the cold volume has reached its maximum value and is charged with cold gas at the full intake pressure. The intake valve is closed at the end of the intake stroke.

The expansion operation is carried out by opening the exhaust valve slowly over a period of time when the displacer is essentially motionless. If the expansion occurs too abruptly, poor heat transfer will result between the exhausting gas and the metal matrix of the regenerator. As expansion occurs the gas remaining in the cold volume falls to a progressively lower temperature. The gas leaving the cold volume is partially warmed by heat removal from the working temperature (the refrigeration load in Fig. 1) and is subsequently heated all the way to room temperature by passage through the regenerator.

The cycle is completed by moving the displacer so as to expel the cold expanded gas remaining in the cold volume at the end of the expansion operation. The exhaust valve remains open as the displacer is lowered and is closed after the exhaust stroke is completed. The initial condition is now restored and the whole sequence of operations is repeated.

THERMODYNAMIC CONSIDERATIONS

Each cycle results in an enthalpy change of $\Delta H = +(p_1-p_2)V$ for the gas at the warm end of the refrigerator and a corresponding change of $\Delta H = -(p_1-p_2)V$ at the cold end, where p_1 and p_2 are the high and low pressures and V is the displaceable volume. This can be demonstrated by reference to Fig. 2 which shows the pressure–volume diagram for a complete cycle for (a) the warm end and (b) the cold end. The two diagrams are identical except for the fact that the p–V area is circumscribed in the counterclockwise direction at the warm end, corresponding to the generation of enthalpy, and in the clockwise direction at the cold end, corresponding to a reduction of enthalpy. Since all other processes involved are constant-

pressure processes, and since there is no mechanical work delivered, it can be concluded that, ideally, an amount of heat $(p_1-p_2)V$ is pumped from the

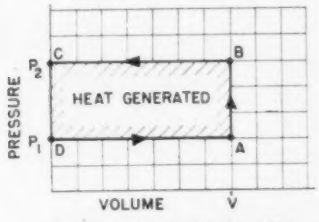


FIGURE 2 (a). WARM VOLUME P-V DIAGRAM

(a)

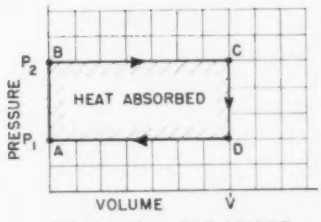


FIGURE 2(b). COLD VOLUME P-V DIAGRAM

(b)

FIG. 2. p–V diagrams (A–B pressurization, B–C intake, C–D expansion, D–A exhaust). (a) Warm volume. (b) Cold volume.

low temperature and rejected at room temperature for each cycle.

There are many ways in which the machine can deviate from ideal behavior. The principal sources of inefficiency can be divided into two groups, thermal inefficiencies and pressure–volume inefficiencies. Thermal inefficiency can result from inadequate heat transfer or insufficient heat-storage capacity in the thermal regenerator or at the site of the refrigeration load. Heat leakage from the warm to the cold ends of the displacer–cylinder assembly or along the regenerator also contribute to thermal inefficiency. On the other hand, poorly arranged valve timing or restricted flow passages may seriously diminish the actual pressure–volume area, thereby reducing the enthalpy change per cycle. Finally, unnecessarily large gas volumes in the flow passages or in the regenerator, valves or connecting tubes result in gas wastage, because the static volume fills with gas and discharges each cycle without contributing to enthalpy change.

Generally speaking, minimization of thermal inefficiencies is incompatible with minimization of pressure–volume inefficiencies. For example, an ideal thermal regenerator would have extremely small flow passages, a large surface area for good heat transfer and a large volume of metal for good heat storage. This is clearly incompatible with the requirement of minimum internal gas volume and minimum restriction to gas flow. In order to achieve maximum efficiency for the refrigerator, it is necessary to optimize the design in multi-dimensional space, taking into account at least ten independent parameters. These parameters are: length and diameter of regenerator, density, specific heat and degree of subdivision of metal matrix, length and frequency of stroke, length and diameter of piston, and maximum and minimum pressure. Some of these problems have been considered in more detail in a previous publication.[4] An actual single-stage refrigerator shows an efficiency which is dependent on the temperature span, the efficiency decreasing as the span is increased. Several refrigerators have been constructed which produce 65–70 per cent of $(p_1-p_2)V$ in the form of useful refrigeration, as measured by an electrical heater, at approximately 100°K, where p_1 and p_2 are 300 and 75 lb/in^2 above atmospheric pressure.

THE MULTI-STAGE CYCLE

It has been noted that the efficiency of a single-stage refrigerator decreases as the temperature span increases. This suggests the possibility of using several stages arranged end-to-end in order to minimize the span covered by any single stage and thus permit the attainment of very low temperatures with high efficiency. Fortunately, the present cycle lends itself very well to this arrangement, as shown schematically in Fig. 3. The operating cycle in this case is exactly the same as for the single-stage machine. Refrigeration is produced at three different temperatures: 80°K, 35°K and 14°K in the example shown. The relative magnitudes of the displaced volumes determine the amount of refrigeration produced at each

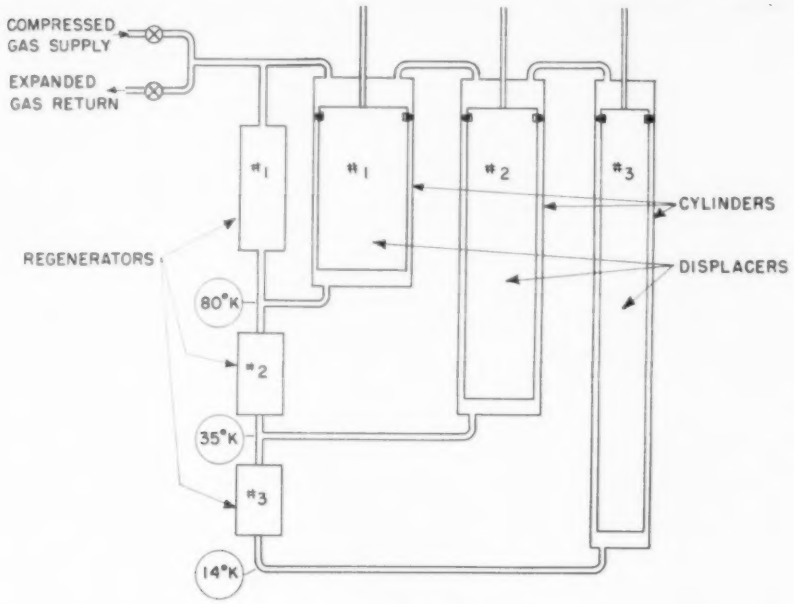


FIG. 3. Three-stage refrigerator.

temperature and will therefore vary from one application to another.

The lowest attainable temperature at the cold end of the third stage is determined by the performance of the third-stage regenerator. The heat capacity per unit volume of all metals diminishes very rapidly at temperatures below 20°K and their heat storage ability becomes poor. The difficulty is further aggravated by the fact that the heat capacity per unit volume of helium gas increases in inverse proportion to the temperature. As a practical matter it has been found that 13–14°K may be regarded as a lower limit if reasonable efficiency is to be preserved, although it is possible to achieve temperatures of 6–8°K if no heat load is applied.

In the multi-stage cycle, refrigeration produced at intermediate temperature levels is used for three general purposes. Firstly, it is used to make up for the inefficiency of the thermal regenerator which feeds gas to that particular temperature level. Secondly, it is used to intercept heat which is flowing into the colder zones, either by conduction along the displacers and cylinders or by radiation from the warm enclosure. And finally, in the event the machine is being used as a liquefier, refrigeration at the intermediate temperatures is used to cool down that portion of gas which is actually

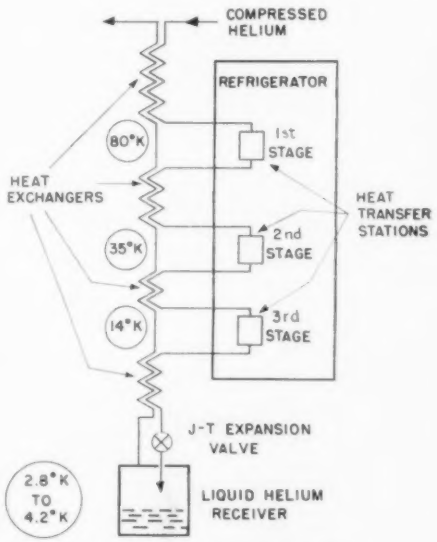


FIG. 4. Helium liquefaction cycle.

liquefied so as to minimize the refrigeration load on the final stage.

Heat transfer between the thermal load (regardless of its origin) and the refrigeration source is established by means of heat-transfer stations which form a part of the flow channels connecting the regenerators to the expansion chambers at the cold end of each cylinder. Such a heat exchanger is shown in Fig. 1, but they are omitted from Fig. 3 for simplicity.

THE HELIUM LIQUEFACTION CYCLE

A three-stage refrigerator such as is shown diagrammatically in Fig. 3 can be modified by the addition of a simple Joule–Thomson cycle to achieve helium liquefaction as shown schematically in Fig. 4. The Joule–Thomson stream, flowing continuously, is cooled by countercurrent flow with the unliquefied portion of gas and also by bringing it into contact with the heat-transfer stations of the three-stage refrigerator at 80°K, 35°K and 14°K. If helium gas at 300 lb/in^2 pressure is cooled to 14°K and expanded to 1 atm in a simple Joule–Thomson cycle, approximately 10 per cent of the gas is liquefied; the remaining 90 per cent is

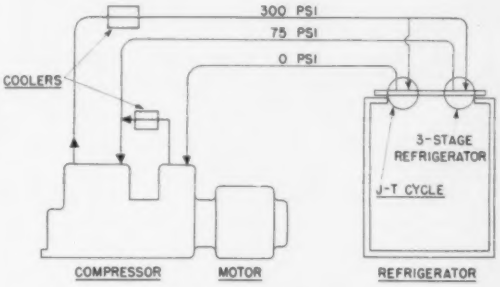


FIG. 5. Helium circulation system.

Table 1. Description of helium refrigerator

	Model no. 1	Model no. 2
*Compressor**:		
J–T cylinder intake	O p.s.i.g.	O p.s.i.g.
J–T cylinder exhaust	75 p.s.i.g.	75 p.s.i.g.
Refrigerator cylinder suction	75 p.s.i.g.	75 p.s.i.g.
Refrigerator cylinder exhaust	300 p.s.i.g.	300 p.s.i.g.
Volumetric displacement J–T cylinder	7·63 ft^3/min	3·85 ft^3/min
Volumetric displacement refrigerator cylinder	15·3 ft^3/min	3·85 ft^3/min
Power consumption	10 h.p.	3 h.p.
Weight: motor and compressor	359 lb	183 lb
Size: motor and compressor	28 × 17$\frac{1}{8}$ × 14 in.	20 × 12$\frac{1}{2}$ × 12 in.
Refrigerator:		
Electrical watts at 14°K	15 W	5 W
Electrical watts at 4·2°K	4·0 W	0·750 W
Electrical watts at 3·2°K	3·0 W	0·220 W
Electrical watts at 2·8°K	0·5 W	—
Size of refrigerated space	1 ft^3	0·1 ft^3
Size of refrigerator	5 × 2 × 2 ft	27 × 15 × 9 ft
Weight of refrigerator	500 lb	90 lb
Length of no. 1 cylinder	10 in.	7 in.
Length of no. 2 cylinder	20 in.	12 in.
Length of no. 3 cylinder	30 in.	17 in.
Diameter of no. 1 cylinder	1$\frac{3}{4}$ in.	1$\frac{1}{4}$ in.
Diameter of no. 2 cylinder	1$\frac{1}{2}$ in.	$\frac{1}{2}$ in.
Diameter of no. 3 cylinder	$\frac{7}{8}$ in.	$\frac{7}{16}$ in.
Stroke	1 in.	1 in.
Rev/min	75 rev/min	105 rev/min

* Model no. 1 water-cooled; no. 2 air-cooled.

recirculated and recompressed. The over-all recirculation system is shown in Fig. 5.

Several closed-cycle refrigeration systems have been built to date which have operated successfully at temperatures of 4·2°K and lower. While each of these experimental machines has differed from the others in some respects, they have all had certain characteristics in common. In each case the system was closed and indifferent to atmospheric pressure. The three operating pressures are established by pressure regulators which either draw gas from or deliver gas to a small storage tank permanently connected to the system. In all cases the compressors have been standard commercial Freon compressors somewhat modified for compression of helium. Special provisions have had to be made for adequate lubrication and oil separation from the compressed helium. Figs. 6–8 are photographs of the refrigerator, the refrigerator in its enclosure, and the compressor system, respectively, referred to as "Model no. 1" in the preceding table. Figs. 9 and 10 are photographs of "Model no. 2" refrigerator and its air-cooled compressor system.

Operating experience to date has been limited to fairly short running periods of 50 hr or less because of the developmental nature of the present program. Nevertheless, the accumulated experience leads to conviction that the reliability of these machines can be excellent. As in all mechanical equipment, the final judgment regarding reliability cannot be made until many thousands of hours of experience have been accumulated.

SUMMARY

A new closed-cycle helium refrigerator is described which essentially eliminates cold moving parts and therefore appears capable of unusual simplicity and high reliability.

It is expected that refrigerators utilizing these design principles will eventually firmly establish the feasibility of superconductive circuits and other practical applications of liquid-helium temperatures.

Present refrigerators require approximately 3 kW for each watt of energy dissipated at 4·2°K; however, it is expected that this figure can be reduced by at least 50 per cent.

Acknowledgement—Much of the work reported in this paper was supported by the International Business Machines Corporation. This support, and permission to publish, is gratefully acknowledged.

REFERENCES

1. H. K. Onnes, *Commun. Kamerlingh Onnes Lab. Leiden* 108 Suppl. 21a (1908).
2. S. C. Collins, *Rev. Sci. Instrum.* **18,** 157 (1947).
3. A. Latham, Jr. and H. O. McMahon, *Refrig. Engng.* **57,** 549 (1949).
4. H. O. McMahon and W. E. Gifford, *Proceedings of the Cryogenic Engineering Conference, Berkeley, Calif., September,* 1959.
5. W. E. Gifford and H. O. McMahon, *Proceedings of the 10th International Congress of Refrigeration, Copenhagen, Denmark, August,* 1959.

FIG. 6. Internal structure of model no. 1 refrigerator.

[*facing p.* 278

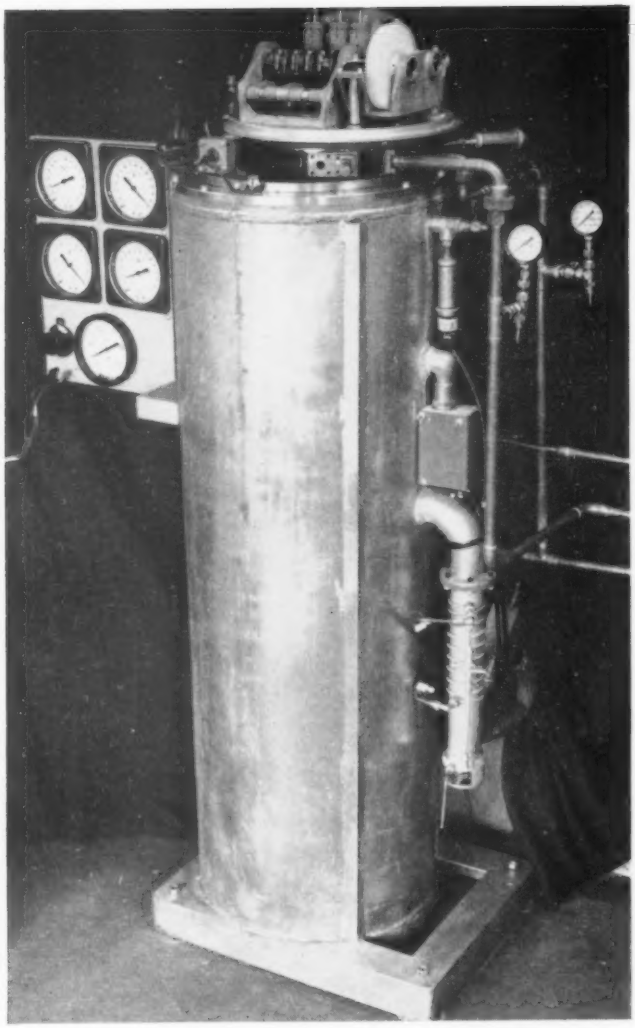


FIG. 7. Model no. 1 closed-cycle helium refrigerator.

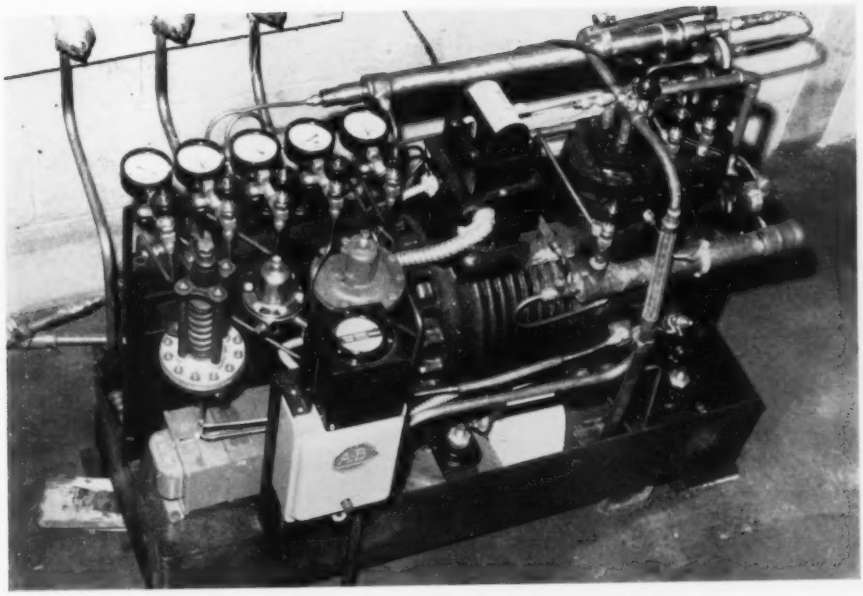


FIG. 8. Compressor assembly for model no. 1 refrigerator.

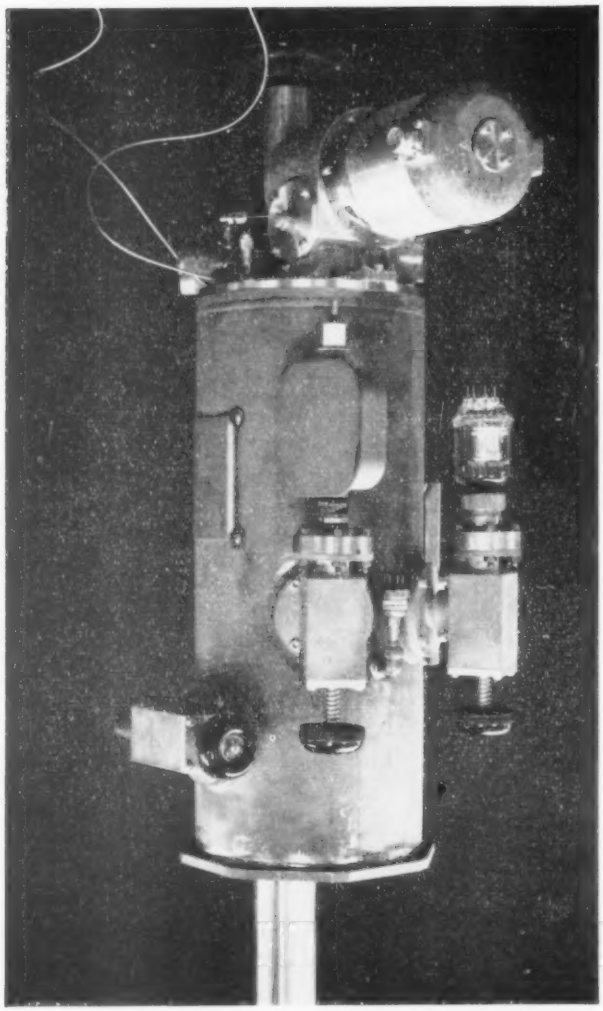


FIG. 9. Model no. 2 close-cycle helium refrigerator.

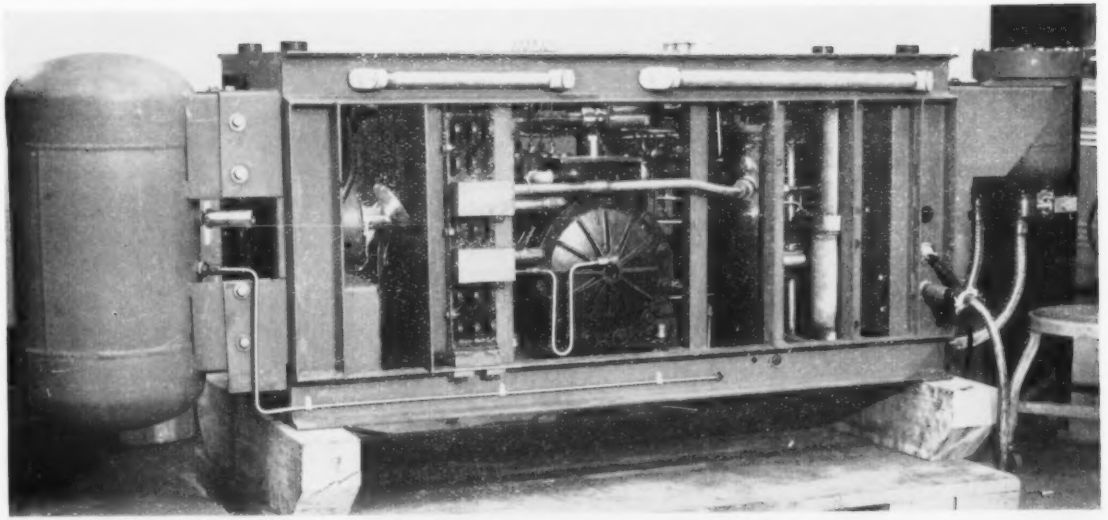


FIG. 10. Compressor assembly for model no. 2 refrigerator.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 279–286. Printed in Great Britain

AN APPROACH TO THE EXPERIMENTAL STUDY OF PERSISTENT-CURRENT DEVICES

C. R. VAIL, M. S. P. LUCAS, H. A. OWEN and W. C. STEWART

Duke University, Durham, North Carolina

Abstract—As part of an investigation of superconducting circuitry, with particular reference to the persistatron, there is being conducted an experimental study of the effects of composition and geometry on the property of superconductivity in thin-film strips. Also under way are an experimental study and a preliminary analysis of the static and transient characteristics of the persistatron. Certain techniques employed in these studies are discussed. Among these are: the evaporation of vanadium and tantalum by electron bombardment, and of lead by the boat method; methods of masking, with reference to both metal and printed masks; and a method of testing persistent-current devices directly in a helium storage flask. The description of a flexible programmed-pulsing device is also given. Results obtained to date in these studies are presented.

THE PERSISTATRON

In a paper presented in August 1957 before the Low Temperature Physics Conference at Madison, Wisconsin, M. J. Buckingham of the Physics Department at Duke University described for the first time a persistent-current device which he called the *persistatron*. Fig. 1 shows the basic persistatron as a ring of superconducting material with a pair of terminals so arranged as to provide two parallel paths of different length—and therefore different inductance—between terminals.

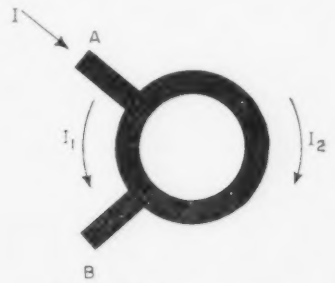


Fig. 1. The persistatron.

If an input current I is fed to the ring with both branches in the superconducting state, flux linking the ring cannot change, and the current I divides into components I_1 and I_2 in such a way as to insure this. The linkage flux ϕ produced by the currents is

$$\phi = I_2L_2 - I_1L_1 \tag{1}$$

If ϕ is initially zero, the current divides so that

$$I_1/I_2 = L_2/L_1 \tag{2}$$

If $L_2 \gg L_1$, then $I_1 \gg I_2$.

If, now, I is increased, I_1 will eventually reach the critical value I_{1c}, and resistance will begin to appear in branch 1. Since branch 1 is no longer superconducting, it is now possible for the linkage to change, and consequently the ratio I_1/I_2 is no longer bound by equation (2), but can change. Thus, all of any further increase in I passes through branch 2, while I_1 remains constant at the value I_{1c}. Recalling that linkage flux was initially zero, we can write the following expression for the present value of linkage flux:

$$\begin{aligned}\phi &= (I - I_{1c})L_1 - I_{1c} \cdot L_1 \\ &= IL_2 - I_{1c}L\end{aligned} \tag{3}$$

where: $L = L_1 + L_2$.

If the current I is now decreased, not only I_2, but also I_1, decreases, and the ring is again superconducting, with the linkage flux ϕ locked in.

The foregoing principle leads to the hysteresis diagram that is shown in Fig. 2. Between the parallel sloping lines, the entire ring is superconducting. Along the right-hand sloping line,

branch 1 has a critical current in one direction; along the left-hand sloping line, branch 1 has a critical current in the opposite direction. From

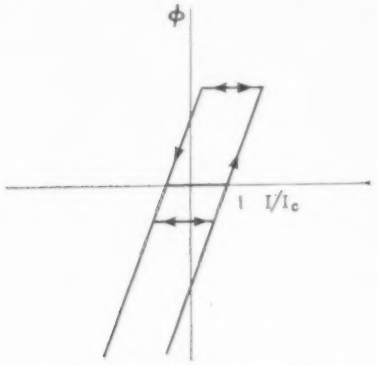


FIG. 2. Persistatron hysteresis loop.

equation (3) it is readily seen that the critical value I_c for the total current I is given by

$$I_c = I_{1c}L/L_2 \qquad (4)$$

It is significant that energy dissipation by the persistatron occurs only in branch 1 and only during periods of change in the value of linkage flux.

THE p-SWITCH

The possibility of using the persistatron as the *memory element* of a computer is obvious. It is also easy to demonstrate the possibility of using the persistatron as a rather unique *switching element*, which can be called the persistatron-switch, or *p-switch*.

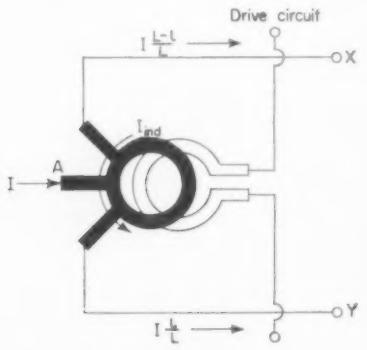


FIG. 3. The p-switch.

Fig. 3 shows a persistatron with three terminal connections instead of only two, together with an inductive loop which serves as a control drive. Suppose that a current in the inductive loop induces in the ring in a counterclockwise direction a current equal to the critical current. A *positive* current entering terminal A will not divide equally between X and Y, but will go preferentially to X. The reason for this is that a current going to Y would combine with the counterclockwise induced current in a segment of the ring to exceed the critical value, thus introducing resistance into the path, whereas the current going to Y would encounter zero resistance. If the output impedance at X and Y is small, say equivalent to an inductance l, the fraction of the incoming positive current that would go to Y would be l/L, while the fraction going to X would be $(L-l)/L$, where L is the inductance of the persistatron. If the incoming current is negative, it would be switched instead to Y.

EXPERIMENTAL APPROACH

Following pioneering of the basic persistatron concepts and of certain of the experimental procedures by M. J. BUCKINGHAM, W. M. FAIRBANK and M. E. BLEVINS of the Physics Department, the authors of this paper, all members of the staff of the Department of Electrical Engineering at Duke University, have continued the work.

From the start of the program, the truth expressed by the following statement of RHODERICK[1] was realized: "It is clear that the future of superconductive switching and memory devices is linked very closely with thin films, from the point of view of both speed and ease of mass production. As far as switching devices of the cryotron type are concerned, the use of evaporated films is in a very preliminary stage and no prediction of their potentialities in this field can possibly be made at the moment. The American approach of undertaking a thorough study of the mechanism by which evaporated films are formed and of their physical properties, before attempting any serious device work, seems essentially sound. Without this there can be no real understanding of the mode of operation of any device, nor can the problem of reproducibility be tackled."

The study of the persistatron as a device has, therefore, begun as a study of thin films. Although

this study has been under way only a relatively short time, the experience already gained may prove useful to others.

FILM FORMATION BY EVAPORATION

Two main methods of evaporation are now being used for the production of experimental thin films. These are: (1) evaporation from a filament of the superconducting material within an evacuated metal vessel which may be either maintained at room temperature or immersed in liquid nitrogen; (2) evaporation in an evacuated vessel at room temperature by electron bombardment. (Also tried was a third method in which the metal is evaporated from a filament or boat onto a substrate which is maintained at the temperature of liquid helium. This method has been discontinued, chiefly because of the time-consuming nature of the procedure.)

(1) *Evaporation from metal filament in metal vessel*

In the earlier work on this project, films, except for those deposited at the temperature of liquid helium, were prepared within a vacuum system employing an oil-diffusion pump and incorporating a glass bell jar and a silicone-greased neoprene seal. Wide variation was noted in the characteristics of films thus produced.

In an attempt to obtain, from consecutive evaporations, films having nearly identical characteristics, an "all-metal" evaporation system was devised. In this system, a metal evaporation vessel is evacuated by means of a mercury pump which is equipped with a liquid-nitrogen trap and which works into a rotary fore-pump. Evaporations can be made either at room temperature or, with the vessel immersed in liquid nitrogen, at much lower temperatures. Fig. 4 shows a photograph of this system.

The evaporation vessel consists of a 7×2 in. o.d. stainless-steel tube to the ends of which $\frac{1}{4} \times \frac{1}{4}$ in. brass flanges have been silver-soldered. The top flange bolts to a mating flange in the lead from the vacuum pump. Details of the lower flange are shown in Fig. 5. It will be observed that the substrate mount is secured to two vertical rods in such a manner as to permit its height above the filament to be altered by a simple adjusting device. Electrical connections to the filament are made via two Kovar seals.

(2) *Evaporation by electron bombardment*

Interest in the technique of electron bombardment has been stimulated recently through its application in floating-zone refining,[2] in the controlled deposition of alloy films,[3] and in the continuous evaporation of refractory metals used in ion-type vacuum pumps.[4] HOLLAND[5] described various methods by which an electron beam may be used to evaporate a metal contained within a crucible. KELLY,[6] in preparing zirconium and tantalum films, first used the pendant-drop method of electron bombardment, which is described below.

Use of this method on the present project followed several unsuccessful attempts to evaporate tantalum, niobium and vanadium by other methods. (The relatively high superconducting transition temperatures of these three metals renders them particularly attractive in this work.)

The operation of the electron-bombardment evaporator may be compared with that of a typical temperature-limited vacuum diode. Each has an evacuated enclosure, a filament or cathode, and an anode. In each, an electrostatic field is maintained between the cathode and the anode, and electrons, upon release from the cathode by

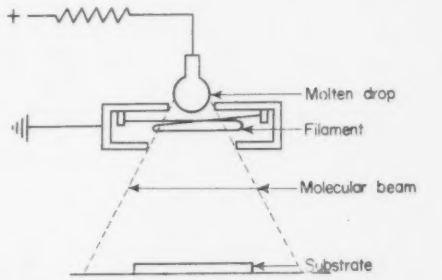


FIG. 6. Electrode and substrate layout.

electrical heating of the filament, are accelerated toward the anode by this field. In the normal vacuum diode the anode design is such as to minimize heating by the impacting electrons, whereas in the electron-bombardment evaporator, the anode, which is composed of the material to be evaporated, is allowed to overheat to the point of melting.

In the pendant-drop method of evaporation, the rod of metal to be evaporated, the anode, is

suspended vertically above the cathode. Provision is made for vertical and lateral adjustment of the rod. No crucible is required, for the molten metal is supported on the tip of the anode by its own surface tension.

The cathode assembly consists of a copper box formed from two channel-shaped sections mounted on two solid copper bars which serve as filament-supply connection terminals, one of the bars being insulated from the rest of the assembly by a mica plate. A single loop of tungsten wire connected between the two copper bars constitutes the filament. The electron beam is brought to a focus on the anode by a $\frac{3}{8}$ in.-diameter hole punched in the center of the upper section of the copper box. Experience has shown that a hole of this size provides sufficient focusing to cause rapid melting of a $\frac{3}{16}$ in. tantalum rod. The molecular beam from the molten anode reaches the substrate below the cathode assembly through a $\frac{5}{8}$ in. hole punched in the center of the lower section of the box.

To minimize the possibility of contamination of the substrate by the filament, the filament is so shaped and arranged as to be "seen" by neither the anode nor the substrate.

Copper was employed in the construction of the cathode assembly because of its low vapor pressure and high heat conductivity. In the present system, cooling of the assembly takes place by way of heat conduction through the current feed-throughs. An improved system employing water cooling is now being planned.

Cleaning of the cathode assembly is facilitated by its construction of two channel sections and two bars.

Fig. 6 shows the layout of the evaporating system, and Fig. 7 shows details of the cathode assembly.

The filament current is supplied by a Variac-controlled small welding transformer. The high voltage required by the anode is supplied by a Variac-controlled full-wave rectifier having a partially smoothed output and capable of delivering 5 kV at 500 mA. Successful operation is possible only with a current-stabilizing resistor connected in series with the anode. This resistor has a value of 18 kΩ.

The bombarding beam of electrons which constitute the anode current causes a molten drop of metal to appear on the tip of the anode. Evaporation takes place from the surface of this pendant drop. For vanadium, it was found that successful evaporation could be achieved with anode currents ranging between 10 and 16 mA at a nominal anode potential of 5 kV. If the bombardment current is allowed to increase above a certain level, the violent oscillations of the drop which were reported by KELLY[6] are seen, and in extreme cases the drop may attach itself to the cathode focusing plate via a very fine spike, thus effectively concluding the evaporation.

In the evaporation of tantalum, vanadium and niobium it has been observed that, during the period of initial melting, violent outgassing of the anode occurs, with accompanying scattering of small particles of metal about the system. This outgassing does not occur during subsequent evaporations. Initial formation of the molten drop is facilitated if a small groove is cut around the anode at a distance of approximately 0·2 in. from the tip. The groove functions as a thermal resistance to concentrate the heat at the tip.

With the vacuum system available to date, none of the films of tantalum, niobium or vanadium prepared by electron bombardment has been superconducting at 4·2°K. In fact, all such films have acted as temperature-invariant resistors in the range of 4·2°K to room temperature. The absence of superconducting transitions in these films is not surprising when one considers that the best vacuum obtainable in the vacuum vessel could not have exceeded 10^{-4} mm Hg, and that these very metals are employed as "getters". In this connection, RHODERICK[1] reports that tantalum films prepared in a vacuum of approximately 10^{-6} mm Hg exhibited superconducting transitions at a critical temperature of 1·6°K, as compared with 4·4°K for bulk tantalum.

Now under construction is a new vacuum system which will permit the attainment of vacua of the order of 10^{-7} mm Hg. In this system the substrate and the molecular beam will be surrounded by surfaces at the temperature of liquid nitrogen. It is expected that these measures will enable the evaporation of niobium, vanadium and tantalum films which will be useful as superconductors.

MASKING OF THE SUBSTRATE

The deposition of a film of evaporated material

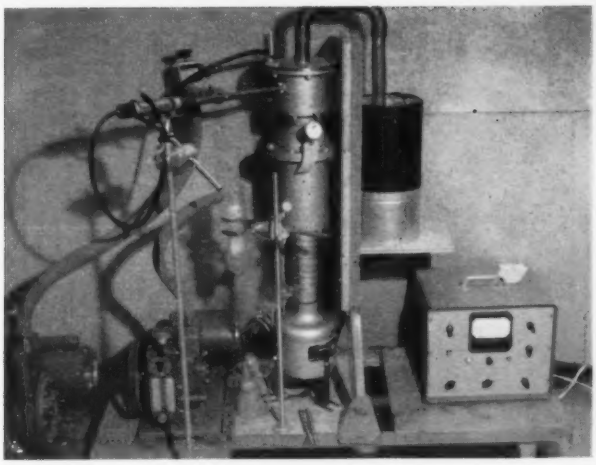


FIG. 4. All-metal evaporation system.

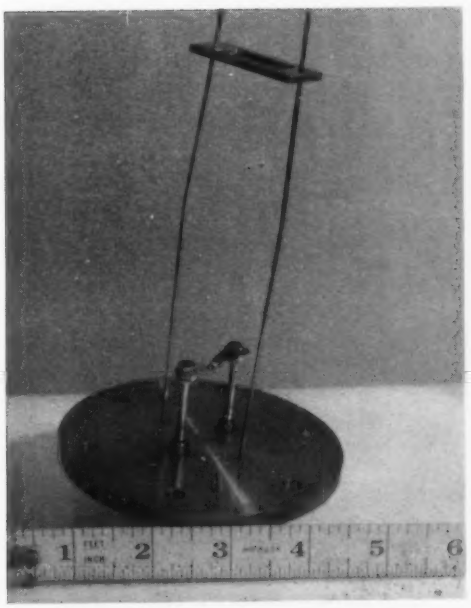


FIG. 5. Lower-flange assembly.

[*facing p.* 282

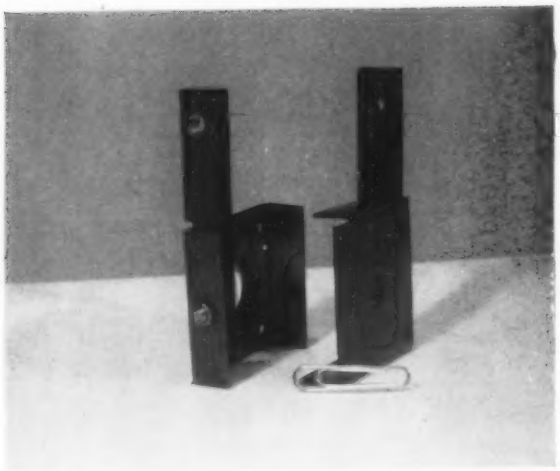


FIG. 7. Cathode assembly.

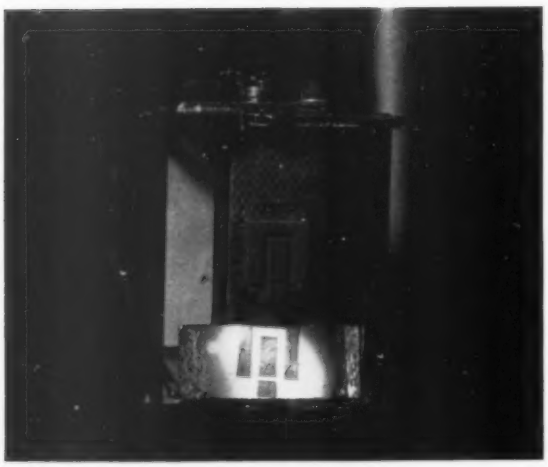


FIG. 8. Rubber stamp and persistatron.

in a particular geometrical pattern, such as that of the persistatron, has been accomplished by means of masking of the substrate. Several different masking techniques have been employed with varying degrees of success.

The geometry of the persistatron is such that, in order to obtain an open interior region in addition to a narrow bridging element across the connection leads, either a two-piece mask or a double evaporation must be employed. Wherever possible it has seemed desirable to use those methods which require only single-stage evaporation in order to reduce the chance of obtaining non-superconducting connections at the junctions of the separately evaporated portions of the circuit. Nevertheless, both single- and double-evaporation methods have been employed and will be discussed.

(1) *Parallel-line mask*

This method of masking, which requires two separate evaporations to obtain the complete persistatron circuit, makes use of a mask of thin metal in which a number of parallel slits have been milled. The thinnest possible metal consistent with fabrication techniques and mounting arrangements is employed. In order to minimize loss of definition in the evaporated pattern, masks are mounted as close as possible to the substrate.

After one set of parallel lines has been placed on the substrate by the first evaporation through one-half of the mask, the mask is rotated 90° on its pivoted mount, thus bringing the other half of the mask into position over the substrate with the milled slits perpendicular to the lines deposited by the first evaporation. A second evaporation now deposits the second set of lines on the substrate, to yield a pattern in the form of a rectangular grid. Upon removal of the substrate from the evaporator, unwanted portions of the grid pattern are disconnected by cutting with a pointed instrument.

An elaboration of this method which would avoid the deposition of unwanted parts of the grid could be accomplished by appropriate design of the slits in the mask.

A fact which may or may not be of significance with respect to the operating characteristics of a persistatron, is that the configurations produced by the parallel-line method are rectangular, rather than circular.

A fact which is significant is that the double evaporation may result in non-superconducting junctions between the two layers. Such a persistatron element would not function.

(2) *Negative mask*

One method of surmounting the difficulties introduced by multiple evaporations makes use of a mask formed as a negative of the persistatron. The exterior portion of the mask is made by punching the thin stock to the desired shape with a small hand-operated bench punch. The interior portion is often simply a thin washer. Several methods of fixing the position of this interior portion have been tried. One simple and effective method employs a spot of silicone vacuum grease on the substrate as an adhesive. Another method that has been used is to make the mask of magnetic material and to hold it in place against the substrate by means of small channel-type permanent magnets mounted behind the substrate. Still another fairly obvious approach to obtaining the central open area is to use a substrate having a pre-drilled hole of appropriate dimension in place of the central open area. This last method requires some care to prevent deposition of a film on the wall of the hole, or at least a careful removal of it from the wall, but is has been used with some success.

(3) *Grease mask*

A technique of masking which has been only slightly explored but which seems to hold promise requires no central mask with its attendant locating and fixing difficulties. Under this technique, the circuit to be reproduced by the deposition of evaporated metal is first drawn as a negative to an enlarged scale (such as four to one). From this drawing, a line cut to the desired actual size is made, and from this cut a rubber stamp is prepared. This stamp, when given a thin coating of silicone grease, can be used to stamp a mask of grease on the substrate.

Application of the grease to the stamp is accomplished by touching the stamp to a sheet of glass on which a very thin coating of vacuum grease has been placed. Such a thin coating may be obtained either by rolling out a small piece of grease, or by flowing onto the glass a solution of grease in trichloroethylene and then allowing the trichloroethylene to evaporate. If sharply defined

edges to the final grease mask are to be obtained, the coating of grease placed on the stamp must be extremely thin.

This method has produced metallic film strips of 0·015 in. width with fairly good edge definition. Fig. 8 shows a rubber stamp and the persistatron produced through its use by this method.

Some deterioration with time of the narrow evaporated films produced by this method has been noticed. At the present time of writing it is not known whether this results from a possible small quantity of grease under the film, due to possible smudging at the time of stamping the grease impression, or whether it results from a possible migration of grease after evacuation of the evaporation vessel. No attempt has been made so far to clean the grease mask from the substrate following deposition of the desired thin film.

ELECTRICAL CONNECTION TO EVAPORATED FILMS

Electrical connection to the thin-film circuits has been accomplished with reasonable success by several different methods. One method employs connection lands of fired silver paint to which leads can be attached with low-melting-point solder. Another method, which can be successfully used with very thick films, is to solder directly to the film. The most commonly used method employs lands of indium which are made by heating a piece of indium with a small soldering iron to the point at which it will wet the glass substrate when rubbed against it.

INSTRUMENTATION

The memory function of the persistatron is conveniently tested by means of current pulses of variable width and amplitude and short time rise. Electronic and transmission-line pulse generators fulfil these requirements adequately. It is also convenient to use varying combinations of positive and negative pulses in rapid sequence for such tests. The many methods of triggering drive-pulse generators in sequence include: step relaying, the use of magnetic core switches, the use of electronic switches, and hand triggering. These techniques involve certain disadvantages, including lack of flexibility, slow speed and high cost.

To overcome these difficulties, it has been found advantageous to develop a switching logic generator that can be used with two pulse generators to provide any combination of positive and negative pulses within a repeating group of ten pulses. This unit is small, flexible and inexpensive, and it eliminates the problem of jitter which is encountered in transmission-line pulse generators.

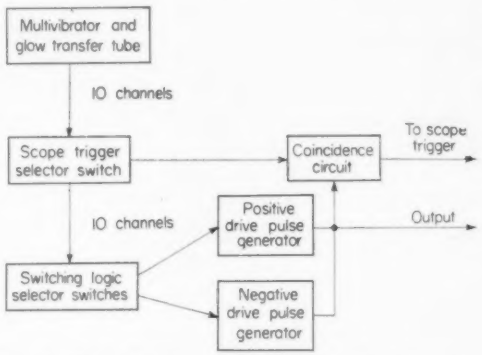


FIG. 9. Block diagram of pulse-generation system.

The principle of operation of this device may be made clear by reference to the block diagram of Fig. 9. A neon glow-transfer tube driven by a multivibrator provides voltage pulses in sequence at each of its ten cathode terminals. A toggle switch in each cathode lead directs the pulse from its cathode to the trigger input of either the positive or the negative drive-pulse generator. A ten-position selector switch allows the signal from any one of the glow tube cathodes to actuate one input of a coincidence circuit. A portion of the output of the drive-pulse generators is fed to the other input of the coincidence circuit, which then provides a signal to trigger an oscilloscope sweep. This feature eliminates the effect of any jitter that may be present between the trigger signal and the output of a drive-pulse generator. It also allows repetitive viewing of the effect of a single drive pulse in the pattern of ten pulses without the superposition of traces from the other nine pulses.

The switching logic generator occupies an 8×8 in. chassis and draws approximately 6 W of power. Pulse repetition frequencies from a few cycles per second up to hundreds of kilocycles per second are obtainable. Drive-pulse patterns are easily altered during operation.

DETERMINATION OF FILM THICKNESS

Most of the films made and tested to date have been of lead. Determination of the thickness of lead films evaporated onto substrates at room temperatures has been made by measurement of the resistivity of the film and reference to a calibration curve, Fig. 10, that relates resistivity and film thickness.

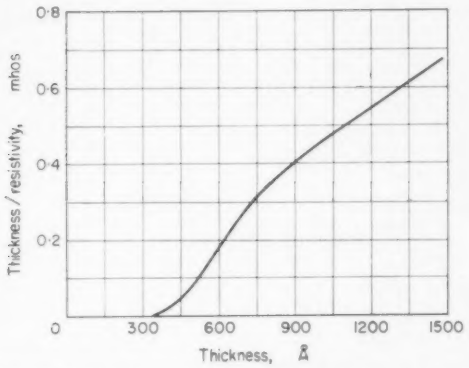


FIG. 10. Resistivity vs. thickness for lead films evaporated on room-temperature substrates.

The calibration curve was obtained experimentally by the evaporation of lead from a large volume source by induction heating. Resistance of the film was then recorded as a function of time, the evaporation continuing until bulk resistivity had been obtained in the sample. With these data and the assumption of constant evaporation rate, resistivity as a direct function of thickness was plotted.

TESTING TECHNIQUES

It has been found advantageous to evaporate superconducting circuits upon substrates less than $\frac{1}{2}$ in. wide so that they could be inserted through the neck of a conventional 15-l. liquid-helium storage dewar directly into the liquid-helium bath. This procedure eliminates the necessity for transferring the liquid helium to another vessel for experimental purposes, and thereby minimizes the helium loss rate.

For testing purposes, the substrate with its specimen is mounted at the end of two small-diameter parallel stainless-steel tubes through which run coaxial cables for current and voltage connections to the specimen. This "dip-stick" has proved to be highly satisfactory. Fig. 11 shows a specimen in place.

TEST RESULTS

The thin-film study so far has been largely concerned with lead strips deposited in moderate vacuums on glass substrates at room temperature. These strips, which have ranged from 500 to 1500 Å in thickness, have exhibited rather non-uniform characteristics. Values of critical current have varied widely for specimens of the same dimensions. When driven by a low-frequency sinusoidal source, many of the strips remained normal throughout the cycle, once the current had attained critical value. Other strips, after having been driven normal, returned to the superconducting state at a lower current level later in the cycle, as Fig. 12 illustrates. The hysteresis loops for films of the latter type remained unchanged over a frequency range from 20 c/s to 20 kc/s. Some strips exhibited multiple transitions that were apparently related to the slope of the edge of the film.

The results of tests on the persistatron are interesting. Fig. 13 shows the volt–ampere characteristic of a persistatron driven by a 3 kc/s sinusoidal current source. In this instance both branches of the superconducting loop were driven normally.

Fig. 14 depicts the response of a persistatron to 1 μ sec current pulses. The memory characteristics of the device were demonstrated in this test.

It has also been noticed that an occasional thin-film persistatron may, when driven with a steady direct current, produce a relaxation oscillation, such as is pictured in Fig. 15. Such an oscillation occurs when current in the shorter parallel branch approaches its transition value. The frequency of the output signal is seen to increase as the value of the input current is increased. When the input reaches the value at which the current in the shorter branch remains normal, the oscillation ceases.

Acknowledgements—The authors of this paper wish to express appreciation to the U.S. Atomic Energy Commission for its sponsorship of this work under Contract no. AT-(40–1)–2430 with the Department of Electrical Engineering of Duke University. Appreciation is also expressed to the Department of Physics of Duke University, whose co-operation has done much to expedite this work.

REFERENCES

1. E. H. RHODERICK, *Brit. J. Appl. Phys.* **10,** 193 (1959).
2. A. CALVERLEY, M. DAVIES and R. F. LEVER, *J. Sci. Instrum.* **34,** 142 (1957).
3. D. D. METZGER, *Process Control for Thin-Film Vacuum Deposition.* Thesis, Massachusetts Institute of Technology (1958).
4. N. MILLERON, Continuous evaporation of refractory metals by electron bombardment without support materials, *Vacuum Symposium Transactions,* 1957. Pergamon Press, London (1958).
5. L. HOLLAND, *Vacuum Deposition of Thin Films* pp. 35–140. John Wiley, New York (1958).
6. J. C. KELLY, *J. Sci. Instrum.* **36,** 89 (1959).

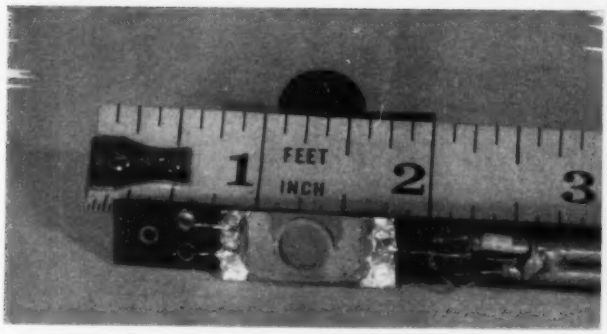


FIG. 11. Persistatron attached to end of dipstick.

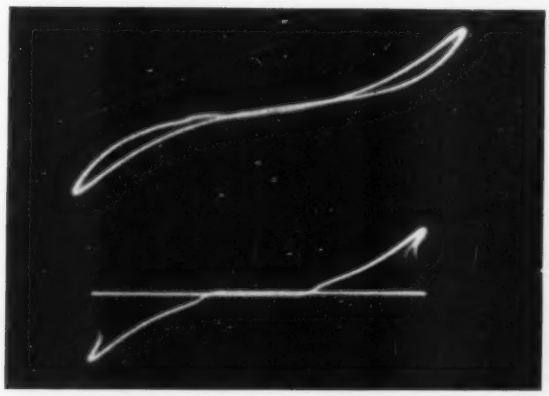


FIG. 12. Oscillogram of voltage vs. current for a lead strip. In upper loop strip is normal over entire circle.

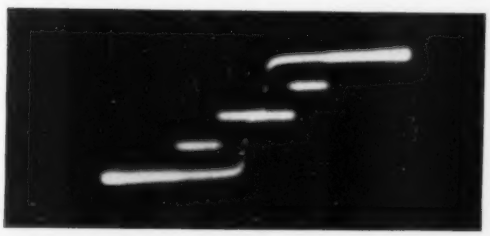


FIG. 13. Voltage output (vertical) vs. current input for a persistatron. Hysteresis due largely to joule heating.

[*facing p.* 286

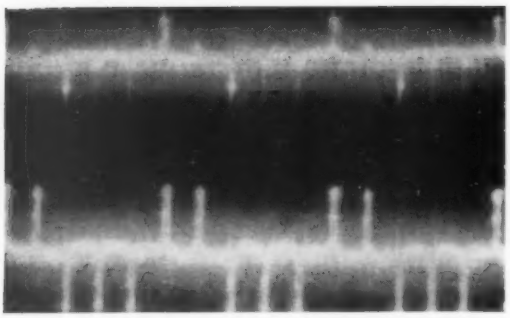


FIG. 14. Persistatron output (upper) for current pulse input (lower). Pulse repetition frequency 16 kc/s.

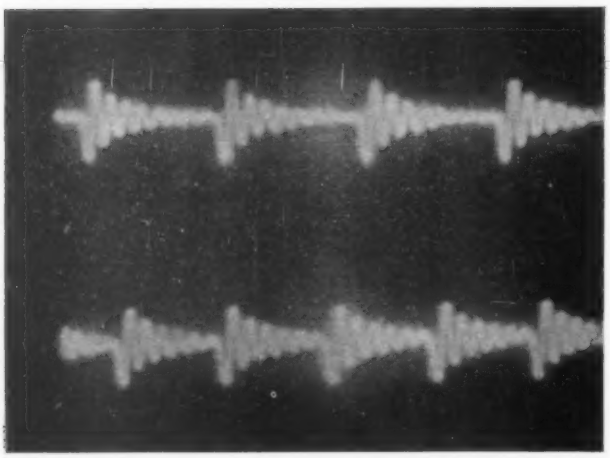


FIG. 15. Two-megacycle relaxation oscillation of persistatron. Input direct current greater in lower oscillogram.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 287–305. Printed in Great Britain

THERMAL AND ELECTRODYNAMIC ASPECTS OF THE SUPERCONDUCTIVE TRANSITION PROCESS*

W. H. CHERRY and J. I. GITTLEMAN

RCA Laboratories, Princeton, N. J.

Abstract—The dynamic processes occurring in the transitions back and forth between the normal and superconducting states dominate the practical aspects and engineering considerations of virtually all superconductive computational devices, yet represent an area of greatest deficiency in theoretical understanding. However, in the case of the superconducting-to-normal transition induced in a thin wire or strip of film by the flow of a sufficient electric current, one of the chief dynamic aspects of the process can be discussed simply on the basis of a proper accounting of all heat generation and heat flow, particularly of the joulean heat generated in an already normal region. The derivations are outlined of the formulas describing the velocity of propagation of the transition and of the relations between current, bath temperature and the quenching characteristics of the material. The approximations necessary are indicated. In many other cases, such as expulsion of or quenching by an external magnetic field, or in certain extreme conditions of current quenching and recovery, the chief aspects of the process are electrodynamic. While the Londons' equations and boundary conditions give a complete characterization of the electrodynamic properties of the superconductor, they do not apply to the transition process itself. Some considerations are given for an appropriate generalization of these equations, and those of more recent theories, to the transition, and a few simple, heuristic models are used to suggest specific equations.

I. INTRODUCTION

WE MAY anticipate for the future a great number of devices and machines which depend in an essential way on the maintenance of the superconducting state and, indeed, which would be rendered inoperative by a transition to the state of normal conduction. However, in most superconductive devices which are intended for use in electronic computers, operation necessarily involves the going back and forth between the normal and the superconducting states. In fact, in most cases the successful performance of the device can be measured by the degree of effectiveness with which these transitions are carried out: with what speed, reproducibility, power-drive requirements, power gain, electrical noise, and so forth. One underlying motivation for the work to be reported here, and it is no doubt a motivation shared by almost all who are working in the field, is to construct a model of both superconduction and this transition process, which will permit the calculation of such figures of merit and circuit characteristics, for given device configuration, and which will suggest how changes in configuration or selection of materials may improve these properties. Needless to say, we can as yet offer no such comprehensive model, even one of a semi-classical nature (as it probably must be for the sake of heuristic and engineering utility). We shall confine ourselves to a discussion of two of the kinetic mechanisms which occur in and, in part, control the transition process. Some aspects of what a workable model must be like will then be inferred from this.

One dynamic process which is of some importance in most types of transitions and in the case of current quenching often dominates the transition behavior, is that attendant upon the generation of joulean heat by current flow in already normal sectors of the material. Another such process, almost always present and often dominating, is the electrodynamic one. Involved here are the relations among field and current and the changing properties of

* The research reported in this paper has been sponsored in whole or in part by the United States Air Force, under Contract AF33(616)–6405.

the superconductive material as an electromagnetic medium. Clearly, there must here be an intimate connection with the Meissner effect and so the latter will enter into the discussion in some detail.

II. THE CURRENT QUENCHING OF SUPERCONDUCTORS

It is a very common experimental situation to pass electric current through a superconducting conductor as part of a circuit, and to find that, as the current is raised beyond some quite reproducible threshold, the superconducting properties are destroyed. Reducing the current will often lead to a recovery of these properties. Since some magnetic field must necessarily be associated with this current flow, and it is known that magnetic fields of moderate intensity are capable of quenching superconductors, it is natural to look first to this mechanism to give a quantitative description of the quenching by current. We shall find, however, that quite apart from the effects of the magnetic field or other mechanisms of quenching by current, the joulean heat generated by the current in already normal portions of the conductor has an extremely important effect on the transition process and the threshold at which it occurs.

In reference to quenching by a magnetic field, PIPPARD[1] proposed a model for the growth of the normal phase at the expense of the superconducting phase. In this model, the interphase boundary moves at just that velocity that the net magnetic field there, from the applied field plus that produced by eddy currents induced in the normal phase by the motion, is equal to the critical field characteristic of the material and its temperature. There, if the growth of the normal region were slower than a certain speed, flux would tend to pile up at the boundary and the field would tend to exceed the critical value; on the other hand, if the growth were faster than this speed, flux would be unable to reach the boundary quickly enough to maintain its critical value. Hence a unique velocity of interphase boundary motion is obtained. The validity of this model has been demonstrated by FABER[2] on cylinders placed in magnetic fields parallel to their axes. However, the presence of eddy currents within the normally conducting material ensures the generation of joulean heat, and ITTNER[3] has discussed the effect of this on the transition process when the applied fields are much greater than the critical field. In our present case, the presence of a primary conduction current offers the possibility of joulean heat generation even without eddy-current effects.

A. Experimental situation to be considered

Consider a superconducting wire of length very large in comparison to its diameter, which is immersed in a liquid-helium bath. It is subjected to a rectangular pulse of current of sufficiently long duration that the processes now to be discussed can reach a condition of dynamic steady state. The initial transient, by which the current is set up along the wire, travels down the wire with the speed of light and is long past before any region of transition from the superconducting to the normally conducting state, which region may be formed initially, can expand or propagate through the wire in any direction whatever.

B. Pippard's transition mechanism

According to the theory of PIPPARD[1], there would be no initial region of transition to the normal state unless the magnitude of the current exceeded that critical value which would produce, at the wire surface, the critical magnetic field of the superconducting material at the given bath temperature (the Silsbee hypothesis). In that event the initial region of transition would comprise the entire cylindrical surface layer of the wire. Then this region, according to PIPPARD, would propagate radially inward with a velocity characteristic of the propagation of eddy-current disturbances in the normally conducting material and with the magnetic field intensity at the moving, normal-to-superconducting interface sustained at the critical field intensity. Once this inward propagation was completed, the wire would, during a further transient, revert to an intermediate state. Because of the smallness of the wire radius, this inward propagation process would, in the majority of experimental conditions, occur much more rapidly than the different process of propagation along the length of the wire which we are about to discuss. For many experimental conditions, this latter process will occur at lower currents than the critical value already mentioned and will dominate that of PIPPARD, in some cases even

precluding the formation of a final intermediate state.

C. *Initiation of the heat-propelled interface transition mechanism*

We postulate the early formation of a region of normally conducting material somewhere along the wire. This may have been brought about before the application of the current pulse, by local heating, a local magnetic field, an additional electrical current passing through one place in the wire, or by an inhomogeneity in metallurgical composition. It may also have been brought about during the rise of the current pulse by a PIPPARD process or some other mechanism operating locally at a metallurgical inhomogeneity, mechanical fault or a constriction in the wire. However, we shall apply our quantitative analysis only to the bulk of the wire which is under both internally and externally homogeneous conditions, and assume that the normal nucleus expands into the homogeneous region by the type of process under study.

D. *Approximations and notation*

For the sake of the great simplification gained in a one-dimensional analysis, we shall assume that it is sufficient to represent the significant variables along the wire, such as temperature, or current density in the normal region, by single mean values averaged over the cross-section. We merely note that such an assumption becomes suspect when, for relatively large wire diameters, the impedance to radial heat flow of the metal increases toward equality to the impedance in series with it presented by the boundary layer between the metal and the liquid helium. Moreover, in the region close to the moving interface between the normal and superconducting parts of the wire, the current distribution shifts from a surface concentration in the superconductor to a uniform distribution in the normal conductor. Extending through this region of shift, there must be large radial variations of current density and consequently more joulean heating than in the bulk of the normal region. Possibly there may also be local regions in the intermediate state, near the interface. All these factors are ignored in the one-dimensional description, as sketched in Fig. 1.

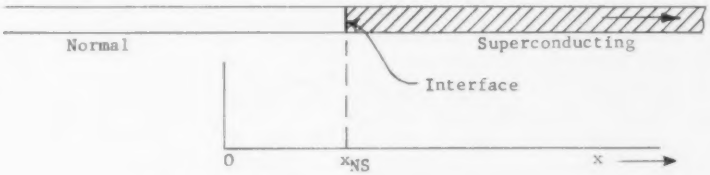


FIG. 1. Idealization of wire under transition.

Let

- x_{NS} = co-ordinate of the normal-to-superconducting interface
- T_t = transition temperature of the material under the given current, and is the temperature at the interface where the transition is presumed to be progressing
- T_C = critical temperature of the material, that is the transition temperature at zero current and zero magnetic field
- T_B = liquid-helium-bath temperature
- $T_N(x, t)$ = temperature of the normal material at point x and time t
- $T_S(x, t)$ = temperature of the superconducting material at x and t
- C_N = heat capacity at constant pressure per unit volume of the normal material
- C_S = heat capacity of the superconducting material
- C_L = latent heat of transition, at constant pressure per unit volume (In the case of thin films, this might be supplemented with an additive term for the surface-energy change)
- k_N = thermal conductivity of the normal material
- k_S = thermal conductivity of the superconducting material
- k_r = thermal conductance (not of the same dimensions as conductivity) in the radial direction of the boundary layer between the metal and the liquid helium
- r = radius of the wire

j = current density in the normal region, assumed uniform over the cross-section
σ = residual electrical conductivity of the normal material

Those of the foregoing quantities which are characteristic of the material may be temperature dependent, current-density dependent and magnetic-field-intensity dependent. In the range of the present experimental work, however, they are constant or their variations can be reasonably well estimated, and C_L is negligible. (The occurrence of a Peltier effect near the interface will be neglected too.) A fact helpful to such approximations is that the phenomenon under study is affected primarily by the values near the interface, and here temperature differences are relatively small. Furthermore, a large amount of information about these quantities can de derived from the data on the propagation phenomenon itself, and so internal consistency checks are quite feasible. However, in the case of thin films, particularly at high current densities, difficulties as to the current dependences of these quantities may be anticipated.

E. Boundary-layer heat conductance

Of particular experimental interest is the boundary layer heat conductance k_r. Except for surface conditions such as polish and state of oxidation, it is probably independent of the metal and, of course, the current and magnetic field.* Its dependence on temperature and temperature difference, and particularly on the change of helium at the λ-point, may be very interesting and of practical value, and since the present propagation measurements lead to a direct evaluation of k_r, this method may be the best means for its study.

F. The heat-balance equations

The heat balance equations for an elementary length of wire at x and t may be written from first principles, thus, term by term: the rate of change of heat content per unit length of wire equals the joule heat generated plus the difference between the heat flowing in along the wire at one end and the heat flowing out at the other end plus the heat loss flowing radially outward to the bath.

$$\pi r^2 C_N \frac{\partial T_N(x,t)}{\partial t} = \pi r^2 \frac{j^2}{\sigma} + \pi r^2 k_N \frac{\partial^2 T_N(x,t)}{\partial x^2} - 2\pi r k_r [T_N(x,t) - T_B]$$

for the normal region, and without the joule heating, the same for the superconducting region

$$\pi r^2 C_S \frac{\partial T_S(x,t)}{\partial t} = \pi r^2 k_S \frac{\partial^2 T_S(x,t)}{\partial x^2} - 2\pi r k_r [T_S(x,t) - T_B]$$

As is the usual procedure with such equations, the desired solution is one which satisfies the differential equation, the boundary conditions and the initial conditions. One such solution having been devised, its uniqueness is usually evident, and in the case of linear systems such as this, assured.

Assuming a very long wire, and considerable time elapsed since the initiation of the current pulse, we want a solution whose character is no longer dependent either on the detailed irregularities at the ends of the wire or on any initial rise-time peculiarities in the current pulse. Thus we want a solution in which such initial conditions have died out and which, though still perhaps time dependent, has a character which is in essence preserved continuously. A traveling waveform would be such a solution. It is of unchanging form but moving with time from left to right. Therefore we introduce the trial solution

$$T_N(x,t) = T_N(\xi), \text{ and } T_S(x,t) = T_S(\xi),$$

where $\xi = x - vt$. (The possibility that v may have negative values must not be overlooked.) Thus

$$0 = k_N \frac{d^2 T_N(\xi)}{d\xi^2} + C_N v \frac{dT_N(\xi)}{d\xi} - \frac{2k_r}{r} T_N(\xi) + \left(\frac{2k_r}{r} T_B + \frac{j^2}{\sigma} \right)$$

and

$$0 = k_S \frac{d^2 T_S(\xi)}{d\xi^2} + C_S v \frac{dT_S(\xi)}{d\xi} - \frac{2k_r}{r} T_S(\xi) + \frac{2k_r}{r} T_B$$

* For possible exception to this statement, see LITTLE[17].

Thus we obtain ordinary differential equations of a very common variety. The solutions are exponentials, and the integration constants are shown as adjusted to the boundary requirements that: (1) at the far left in the normal region, the temperature must represent a balance between joule heat generation and heat loss to the bath:

$$\frac{j^2}{\sigma} = \frac{2k_r}{r}\left[T_N(-\infty) - T_B\right]$$

(2) at the interface, to which we are free to adjust our co-ordinate origin, $\xi = 0$, the temperature is the transition temperature T_t; and (3) at the far right, in the superconducting region, the temperature is the bath temperature T_B.

$$T_N(\xi) = \left(T_B + \frac{r}{2k_r}\frac{j^2}{\sigma}\right) + \left(T_t - T_B - \frac{r}{2k_r}\frac{j^2}{\sigma}\right) \times$$

$$\times \exp\left\{\left[-vC_N + \left(v^2C_N^2 + \frac{8k_rk_N}{r}\right)^{1/2}\right]\frac{\xi}{2k_N}\right\}$$

$$T_S(\xi) = T_B + \left(T_t - T_B\right) \times$$

$$\times \exp\left\{\left[-vC_S - \left(v^2C_S^2 + \frac{8k_rk_S}{r}\right)^{1/2}\right]\frac{\xi}{2k_S}\right\}$$

Fig. 2 shows this temperature distribution along the wire, with a break in slope at the interface to satisfy the heat balance condition there, as explained in the next section.

conditions of heat flow in the neighborhood of the moving interface must be such as to sustain the constant temperature of the interface, since the transition is assumed in the present theory to occur because of heating and hence to take place isothermally. Thus the net heat flow to the interface, from left and right, must be just sufficient to supply the rate of absorption of latent heat of transition. Thus

$$-k_N\frac{\partial T_N(x,t)}{\partial x} + k_S\frac{\partial T_S(x,t)}{\partial x} = vC_L$$

This becomes, on substitution of the differential expressions for $T_N(\xi)$ and $T_S(\xi)$, with $\partial\xi/\partial x = 1$,

$$-k_N\left[T_t - T_B - \frac{r}{2k_r}\frac{j^2}{\sigma}\right] \times$$

$$\times \left[-vC_N + \left(v^2C_N{}^2 + \frac{8k_rk_N}{r}\right)^{1/2}\right]\frac{1}{2k_N} +$$

$$k_S\left[T_t - T_B\right]\left[-vC_S - \left(v^2C_S^2 + \frac{8k_rk_S}{r}\right)^{1/2}\right]\frac{1}{2k_S} =$$

$$= vC_L$$

This equation, then, defines a unique value of the velocity. By measuring this velocity for various values of current density and bath temperature and then using this equation, one can calculate the dependence of transition temperature on current density and also on magnetic field if that is present.

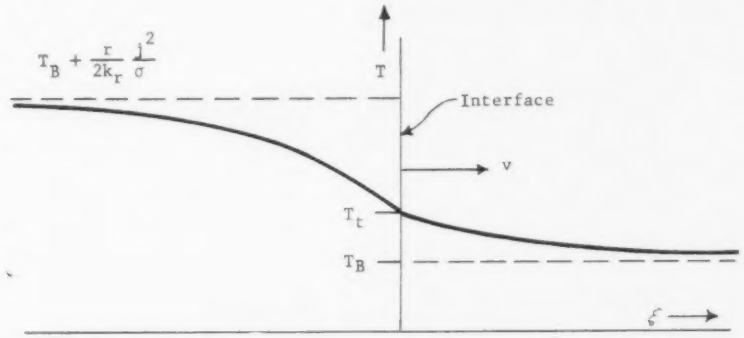


FIG. 2. The temperature waveform moving along the wire.

G. Heat balance at the interface itself

There is one further boundary condition, however, which must be taken into account. The

In order to do this, it is necessary to know the thermal quantities appearing in the equation, or at least know them to be constants within the

range of the experiments; some of their values may be computed from the measurements. If it is experimentally feasible to adjust the current and the bath temperature toward vanishingly small interface velocities, the determination of the dependence of the transition temperature on current density is particularly easy and depends only on the ratio of k_S/k_N.

H. The interface velocity

In the foregoing equation, the explicit solution for v in terms of the other variables proceeds uneventfully, except perhaps for the elimination of the three spurious roots introduced when the equation is expanded to remove the radical signs. The result is a biquadratic equation which can be solved algebraically. However, the ranges of the present experiments were such as to make the latent heat (C_L) term negligible, and hence this term will be omitted from the following for the sake of a considerable algebraic simplification.

With the several thermal quantities characteristic of the material gathered together into four symbols, w, x, y and z, which, at least in some cases, will be unknowns to be determined by these experiments, the velocity is given by

$$v = y\left[(w\beta-1)-\frac{x}{w\beta-1}\right]\times$$

$$\times\left[(x+z^2)+z\left((w\beta-1)+\frac{x}{w\beta-1}\right)\right]^{-1/2}$$

Here, in terms of previously defined quantities,

$$w = 1/(2\pi^2 k_t r^3\sigma), \qquad x = k_S/k_N,$$

$$y = (2k_t k_N/rC_N^2)^{1/2}, \quad z = C_S/C_N$$

while $\beta = I^2/(T_t - T_B)$ encompasses the two principal experimentally controlled variables, the current and the bath temperature.

I. Correlation of the formula with experimental data

If T_t, the transition temperature, is also known, then β becomes fully known and controllable. Assuming that w, x, y and z are constants, their evaluation from measurements of v is relatively simple. A plot of log v vs. log β cuts off both log y and log w as quantities affecting the curve shape because they serve merely to fix the origin of the co-ordinate system. The problem reduces to the selection of that one of a two-parameter (x and z) set of curves which most closely fits the data. Thus all four constants may be evaluated at once, although the curve-fitting method is, because of the shapes involved, rather insensitive and the experimental error large. Even if w, x, y and z are temperature dependent, an extensive range of values for the experimental variables I and T_B should give both the magnitudes and the temperature dependences of these thermal quantities.

The behavior of the equation for v is quite harmonious with intuition. Extremely large values of I, or bath temperature just slightly below the transition temperature, either situation, leads to an extremely large value of β. The joule heating is very high and the velocity of the transition interface is hence also very high. (For bath temperatures in excess of the transition temperature, there is no superconducting state and the equation does not apply.) As β diminishes, so does v, until at $(w\beta - 1)^2 - x = 0$, and hence $v = 0$, the joule heat generated at the left of the interface and just reaching it, is exactly counterbalanced by the heat loss directly to the bath plus the loss through the superconducting region on the right, eventually also to the bath. For still lower values of joule heating, or for even greater cooling by a lower temperature bath, v is negative. This means that the transition is moving from right to left and is one from the normal state to the superconducting state. Such a situation might occur when a higher current, producing the transition moving to the right at an earlier time, is followed by a lower current unable to sustain the transition. This consideration suggests an experimental variation in which a variable current, possibly sinusoidal a.c., may be used to determine the $v = 0$ conditions. For even lower values of β, until $(w\beta-1) = 0$, the negative velocity becomes extremely high, until the bath can overcome the joule heating entirely, without help from the heat flow through the interface. Then the normal region vanishes.

The possible plots of log v vs. log β are thus hyperbola-like, running from a negative vertical asymptote at log $[(1+x^{1/2})/w]$ to a positive, upwardly inclined asymptote of slope 1/2. The sketch in Fig. 3 illustrates these properties, and Fig. 4 is a plot of experimental data against a theoretical curve with parameters $x = 1$, $z = 1$.

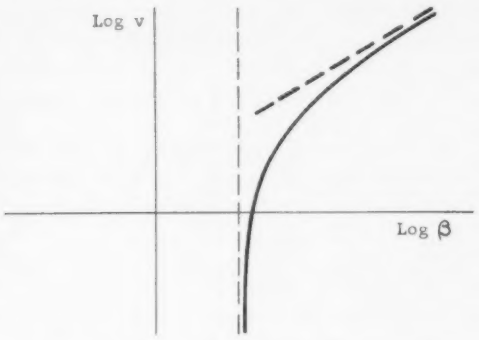


FIG. 3. Sketch of log v vs. log β to show asymptotes.

J. Constant velocity or "isovelocity" plots

Assuming as before that w, x, y and z are constant, which, of course, is reasonably correct over limited ranges of current and temperature, one may collect all of the paired values of current I and bath temperature T_B which lead to the same value of interface velocity. From the formula, which shows v depending only on β, it follows that this collection corresponds to a single value of β; call it β_1. A plot of such I_1^2 versus its paired value of T_{B1} will show directly the dependence of the transition temperature T_t on the square of the current (and also, particularly in the case of thin films, on an imposed magnetic field if that be introduced as an additional experimental variable). Thus $I_1^2 = \beta_1 (T_t - T_{B1})$.

If the transition temperature T_t were T_C, the critical temperature at zero current and field, and entirely unaffected by the current, a set of such plots for different β_1 would obviously be a set of straight lines intersecting the abscissa at a common point: $T_B = T_C$. For T_t, a function of I, the plots, though not in general straight lines, would still intersect or osculate the abscissa at T_C. Thus we have a new method of measuring T_C, entirely independent of conventional ones. The apparently pertinent values of the various thermal characteristics of the material do not actually have to be known.

K. Dependence of transition temperature on current

In bulk materials, and the wires used in the present experiments were bulk under the appropriate criterion which is that the thickness far exceeds the London field penetration depth λ (about 10^{-5} cm), it is known that "Silsbee's hypothesis" applies. This is to say that the dependence of transition temperature T_t on current is precisely that which can be attributed to the magnetic field produced by the current. In the

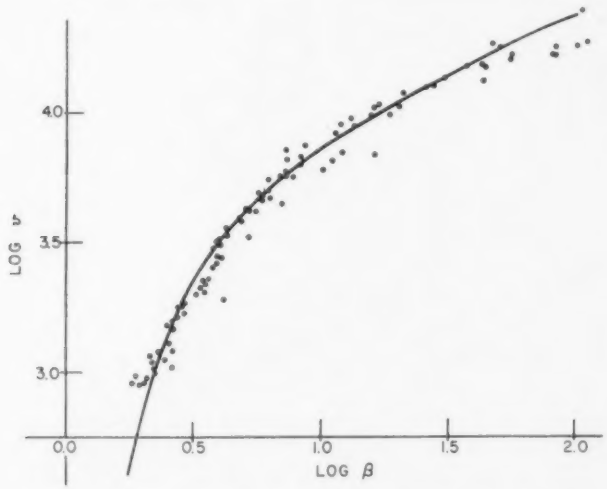


FIG. 4. Plot of interphase boundary velocity v vs. the current–temperature function β: theoretical curve and experimental points for a 0·002-in. tantalum wire.

case of a round wire carrying current, the transition temperature is that defined at its surface where the magnetic field produced by the current is largest. For most materials, the transition temperature is, to good approximation, in parabolic relation to the magnetic field, and hence, in the case of a wire carrying current, also in parabolic relation to the current:

$$T_t = T_C\left(1-\frac{I}{I_C}\right)^{1/2}$$

where I_C is that current which, for the given wire size, produces that magnetic field at the wire surface which will just quench the superconductor at absolute zero temperature. Therefore a plot of the kind discussed in the two preceding paragraphs, showing values of current which, for various bath temperatures, yield the same interface velocity, will have the form of the following equation

$$I_1^2 = \beta_1 T_C\left(\frac{I_1}{I_C}\right)^{1/2} - \beta_1 T_{B1}$$

or

$$T_{B1} = T_C\left(1-\frac{I_1}{I_C}\right)^{1/2} - \frac{I_1^2}{\beta_1}$$

which is common to all the curves. This combination has the effect that, except just at the toes, for all small and moderate values of velocity, and hence also of β_1, the physically significant parts of the curves (down to $T_{B1} = 0$) are nearly linear. Only for very high velocities do the curves display the full extent of the S-shaped term produced by the Silsbee correction. However, caution must be exercised in applying the foregoing formulas to examples of very high interface velocity because the formulas approximate the joule heating in the wire on the basis of a uniform current distribution in the wire cross-section. This becomes inaccurate when the interface velocity approaches or exceeds the inward propagation velocity of eddy-current disturbances. The latter velocity may be estimated from the classic skin-depth formula

$$\delta = \left(\frac{2}{\mu_0\sigma w}\right)^{1/2}$$

showing that the time required to reach the center of the wire is approximately $2\pi/w$ for $\delta = r$; $2\pi/w = \pi\mu_0\sigma r^2$. The velocity is then approximately $1/\pi\mu_0\sigma r$, the dependence on r indicating

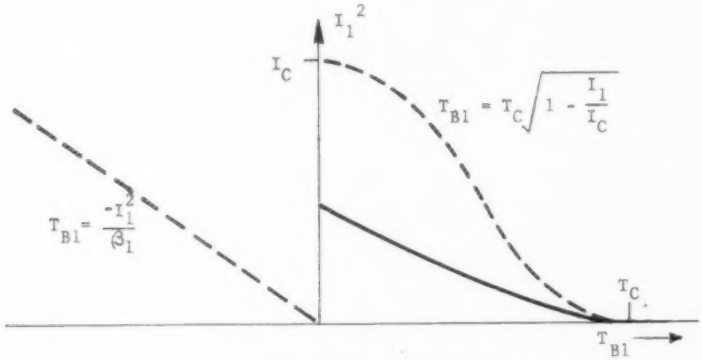


FIG. 5. Sketch of a possible isovelocity plot (solid line).

Any such "isovelocity" curve consists, then, as in Fig. 5, of a linear term ($I_1^2 = -\beta_1 T_B$) with a slope $-\beta_1$, which is different from one curve to another since a change in velocity implies a change in β_1, plus an S-shaped term

$$T_{B1} = T_C\left(1-\frac{I_1}{I_C}\right)^{1/2}$$

that for high interface velocities, present formulas are more accurate for the thinner wires. It will be recalled that the same statement concerning accuracy was implied earlier on the basis of heat-conductivity considerations also.

L. Thin films

While thin films have not been considered

explicitly in the foregoing, it should be apparent that these formulas (except the Silsbee correction which would be quite different) are entirely applicable, provided that factors of the form k_T/r be replaced by k_T/a where a is the thickness of the film, assuming that the film is of rectangular cross-section and exposed to the helium bath on both of its sides. Since in practice the film will be on a solid substrate, the cooling properties of this substrate will have to be included in these factors.

We have obtained a considerable amount of experimental data on tin films which are roughly in harmony with these computations. Fig. 6 is a summary of boundary velocity measurements as a function of current at different temperatures. The observed velocities ranged from 10 to over 100,000 cm/sec. The specimen was a 1-μ tin film, 0·5 mm wide, with a critical temperature of 3·87°K and a residual resistance of 7 mΩ/mm. These measurements were made in the same manner as those for the tantalum wires which, together with a brief outline of the theory just described, were reported by GITTLEMAN and CHERRY[4]. The effects of joulean heat on current quenching in tin films were also described by BREMER and NEWHOUSE[5]. BROOM and RHODERICK[6] have very recently observed similar effects and devised a theory to some extent paralleling the present one.

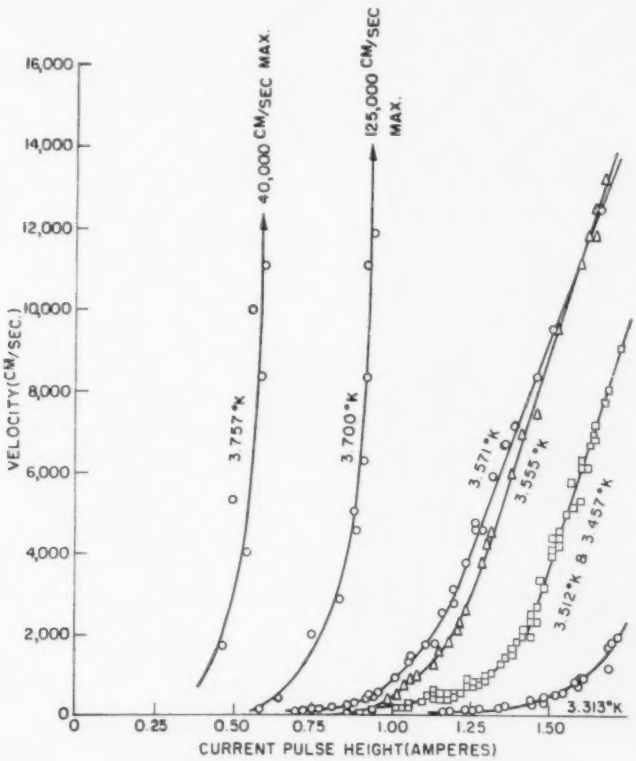


FIG. 6. Interface velocity vs. current for a tin film.

III. ELECTRODYNAMIC ASPECTS OF THE TRANSITION PROCESS*

In the foregoing discussion of current quench-

* One of the authors of this paper (W. H. C.) is the sole author of this section, and wishes to acknowledge much help from consultations with J. I. GITTLEMAN, G. A. MORTON and R. H. PARMENTER.

ing, it was pointed out that electrodynamic processes affect the transition to a quantitatively greater degree when the time needed for electromagnetic field progagation from the conductor surface into its core becomes comparable to the time required for the lateral propagation, through an equal distance, of the quenching process attributable to the current and its associated joulean heating. In many other experimental situations, the electrodynamics are of primary importance under all conditions. When a material going into the superconducting state is used to expel a magnetic field, presumably in accordance with the Meissner effect, or becomes superconducting because a field is withdrawn, the transition process itself is, of course, largely electrodynamic in nature. Even the final equilibrium configuration may depend on the prior electrodynamics. A process of the greatest practical use, since it is essential to most superconductive logic and memory computer elements, is the quenching of a superconductor, perhaps one already carrying current, by an externally imposed magnetic or high-frequency electromagnetic field. The penetration of the field into the superconductor and the reaction of the latter back onto the field are, of course, electrodynamic processes.

In this discussion, after a few, quite evident statements as to what constitutes an adequate and proper specification of the electrodynamic properties of any substance regarded as a medium for an electromagnetic field, we shall examine formulations suitable for a material in the superconducting state, and finally consider what will be required to cover the transition process itself. In regard to this last, one will note that thermodynamic processes undoubtedly accompany the electrodynamic ones. Indeed, depending on the point of view, either may be looked on as part of the other. However, the thermodynamic considerations are suppressed here, partly to achieve simplicity and partly because we do not hope in any case to present more than a beginning to the final answer.

The present considerations will lean heavily on the Londons' formulation[7], in spite of its well-known discrepancies with experiment (penetration depth and high-frequency behavior), for two reasons. Firstly, it can be made a complete electrodynamic formulation, is moderately tractable and, moreover, can be arranged to receive the PIPPARD modification[8] which we take to be supported to some extent by the more fundamental theory of BARDEEN *et al.*[9] Secondly, it is sufficiently appropriate to real superconductors to provide a suitable vehicle for our main conclusions. It rather nearly does describe the principal realities of superconductivity, up to questions about the transition, and present fundamental theory has a shape to conform to it.

A. The specification of electrodynamic properties

The unique, one-to-one correspondence between the complete distribution in space and time of an electromagnetic field and that of the charges and currents which comprise all of its sources implies that either distribution may be calculated from the other. Thus the charge and current distribution may be computed from the field by means of Maxwell's equations, or the field (E and B) may be computed by means of appropriate, retarded integrals over all of the charges and currents. However, if some of this charge and current distribution is contributed through induction by the action of the field on a material medium, neither the field nor the entire charge and current distribution is ordinarily known at the beginning of a computation. In order for the situation to be determinate, there must be specific and unique relationships by which the electromagnetic field induces currents in the medium. Of course, once all of the data from a determinate situation is known, the charges and currents induced in the medium can be computed uniquely from the field by two entirely independent ways, either by the Maxwell equations or by these induction relations. This does not mean that the currents of the medium are overdetermined. Rather, it means that as unknowns they can in effect be eliminated from the problem, much as one might solve simultaneous equations, and the remaining relations permit solving for the remaining unknowns, the field variables. The point of importance is that these induction relations, which in effect characterize the electrodynamic properties of the medium, must convey sufficient information to specify uniquely the induced charges and current in terms of the electromagnetic field entirely independently of any information conveyed by the Maxwell field equations. This

point bears watching since in the literature on superconductivity the field equations and the induction relations are often very much intermingled.

The use of the vector and scalar potentials is of long proven computational advantage in electromagnetic-field calculations. There the indeterminacies, and lack of direct operational definitions, particularly in the case of the vector potential, created no difficulties because the entire procedure could be understood as a matter of algebraic substitution and manipulation, and the indeterminacies or arbitrary determinations disappeared from the final results. The advantage carried over into classical electrodynamics where, either in the Lagrangian or as part of the canonical momentum, the indeterminacies were differentiated away by the time the equations of motion were reached. In further carrying over these potentials, into quantum mechanics, it may be, as has been discussed,[10] that something more fundamental was introduced. While we are inclined toward the negative, this question need not be decided here. It is indeed apparent, however, that in using the vector potential for formulating induction relations, there is a very great chance that one may go far beyond the point where the indeterminacies drop out in the final result. Consider the formulation of the Londons' relations which is often used: $\Lambda J = A$, where Λ is a constant, characteristic of the superconducting material. The current density J is certainly a unique physical entity, yet on the right side of the equation is the vector potential A which, in the absence of additional conditions, is indeterminately defined. In view of the preceding paragraph, it is evident that whatever rule is used for defining A uniquely (choice of gauge), in order to get a definite induction relation for J, conveys part of the information of that induction relation. This rule then must be determined by the nature of the material, and not by any electromagnetic field consideration. But thus we let the material affect the field, or at least an entity representing the field, within an induction relation which is intended to describe how the field affects the material. It appears, therefore, that greater directness, explicitness, and freedom from inadvertent assumptions, are obtained when all induction relations are expressed in terms of the field variables, and the potentials are avoided. An intuitive advantage arises also, in using those agents which act directly, as in the Lorentz force equation, on the charges and currents of the medium, rather than using more indirect conceptual constructs. This matter is touched upon again in Section III C.

B. The Londons' equations

A brief review of the Londons' phenomenological formulation is presented, with some revisions and additions to secure greater logical precision and completeness.

We recall the early discussion of BECKER *et al.*[11] that a free electron fluid, which they supposed is contained in a superconductor, would be accelerated when under the action of an electric field, and the corresponding change in velocity of the charge carriers would appear locally as a change in current density. Thus

$$\frac{\partial J_S}{\partial t} = \Lambda^{-1} E$$

where J_S is the superconducting current density and Λ^{-1} would be proportional to the number of carriers, their charge, and charge-to-mass ratio. It is interesting to note that there is an error here, in equating the rate of change of momentum to the rate of change of current density, because the latter is also in part the result of current carriers moving into the given volume element after having received momentum elsewhere. Hence possible phenomena attendant upon confluence, overtaking and bunching, are overlooked. This equation also omits the Lorentz force acting on the carriers from magnetic fields. Nevertheless, the equation gives a moderately satisfactory phenomenological description of the perfect conductivity aspects of superconductivity. It does not account for the Meissner–Ochsenfeld effect, that is the sudden appearance of almost perfect diamagnetism upon the instant following the transition into the superconducting state. Instead it would predict the "freezing in" of magnetic flux.

From a phenomenological point of view the Londons[7] made up for this lack by postulating the occurrence of an additional pattern of superconducting current immediately after the transition of a material into superconduction. This new current pattern appears as an integration constant in the foregoing equation, integrated from t_t, the

time of transition

$$J_S(x, t) = \Lambda^{-1} \int_{t_t}^{t} E(x, t)\, dt + J_{SB}(x, t_t)$$

(The argument x stands for all three space coordinates.) This integration constant is defined by postulate in part by the equation

$$\nabla \times J_{SB}(x, t_t) = -\Lambda^{-1} B(x, t_t)$$

In this last equation is introduced the idea of a vortex motion of the superconducting fluid in a diamagnetic direction, induced through the interaction of the magnetic field with some mechanism active in the material during the transition. There is, of course, no necessity for a phenomenological description to explain the details of such a mechanism. What is important, however, is that the mechanism is active only during the transition, and although the supercurrents it generates will persist afterward, once the transition is over the mechanism itself is gone, and no new diamagnetic currents are produced by it, even though the magnetic field should change; at least this statement is true, as we shall develop further in the next section, if the Londons' equations give a fair phenomenological description of reality.

The foregoing curl equation is by no means sufficient to define $J_{SB}(x,t_t)$ uniquely, yet uniqueness is necessary for all complete induction relations. There is a vector identity of considerable generality which is derivable from STRATTON's vector form[12] of Green's theorem, which states that, for any analytic vector A,

$$\begin{aligned} 4\pi A = & \int_V (\nabla \times A) \times \nabla(1/r)\, dV + \\ & \int_V (\nabla \cdot A) \nabla(1/r)\, dV - \\ & \int_S (A \cdot n) \nabla(1/r)\, ds - \\ & \int_S (n \times A) \times \nabla(1/r)\, dS \end{aligned}$$

where V is a closed volume, S its surface and r the usual distance from source point to field point. As can be seen in connection with the derivation of this identity, although the details are not particularly germane to the present discussion, this quite general relation is not fully explicit in that it presumes some restriction on the choice of the kernels of the integrals on the right. Subject always to the condition that the curl is divergenceless, only three, but not all four, may be chosen arbitrarily. The specification of the fourth one is then implied (incompletely so in the case of multiply connected regions) by the identity. Its actual elucidation may be quite difficult. The Dirichlet and Neuman problems are special cases of this elucidation. To provide here the additional conditions needed to determine these kernels on the right, we follow the line of thought that the diamagnetic currents are made up of tiny closed loops of current that were generated in the transition. Since such closed loops individually could not exhibit a divergence of the current in them, it is natural to postulate for present purposes that J_{SB} is divergenceless. Moreover, these diamagnetic currents are presumably permanent and hence could not exist while passing through dissipative regions. Thus on the surface of a superconducting region, between the superconducting metal and normally conducting metal, or a non-metal, the normal component of J_{SB} must be zero, and the tangential components of J_{SB} must drop discontinuously to zero outside the superconductor.

Now the foregoing considerations are insufficient to define J_{SB} uniquely in the event that the superconducting region is multiply connected. They do show (from the divergence and normal surface component conditions) that for a singly connected region the surface integral of the normal component of J_{SB} over any simply connected cross-section is zero, so if we further postulate that this is also true for multiply connected regions, we achieve the effect of preventing J_{SB} from coupling or linking around any hole, and thus also achieve the intuitively satisfying result of making the induced superconducting current independent, either as cause or effect, from any magnetic flux passing through a hole in the material and hence entirely remote from the material itself and therefore not acting on it. Thus a material first threaded, through a hole only, by a flux, and later made superconducting, will contain no current produced

by that flux. The electric field induced by a change of such flux is, of course, a different matter and has already been taken into account.

Finally, we consider the normal current-density distribution $J_N(x,t_t-)$ existing in the material in the instant just prior to its becoming superconducting. Intuitively, it seems that the average momentum and direction of drift which the normally conducting electrons must have acquired in order to carry this current must be preserved momentarily through the transition to the superconducting state. Vaguely speaking, if this transition involves merely the removal of hindrances to motion, the pre-existing motion would not be diminished. It seems unlikely that the conduction current flowing just prior to the transition will be instantaneously interrupted in part or in whole when the transition occurs. Of course, physical mechanisms can be imagined which would accomplish this but they are probably not present. Moreover, if such current were interrupted, electromagnetic forces would come into play which would tend to restore it, and in the superconducting medium the resulting transient would disappear rather quickly. Nevertheless, a transient of an experimentally detectable sort would occur and, since this has probably never been observed, it appears safer to assume that it does not occur. (It is somewhat of an extension of the spirit of the Londons' equations, which ordinarily confine themselves solely to the superconducting state, to consider the situation across the transition, but it is clearly necessary to do so in order to make the initial conditions both definite and knowable.) Therefore, we postulate that a term be added to J_S, namely, $fJ_N(x,t_t-)$ where f is that fraction of the normally conducting electrons which becomes superconducting in the given transition, to use the language of the two-fluid model. It may be conceded that from a logical point of view, one could just as easily postulate that this term is zero. (Some intermediate function would appear even more arbitrary.)

Ohm's law still applies to those electrons which remain normally conducting, again using the two-fluid description, and so the superconducting current is always supplemented by $J_N(x,t) = \sigma E(x,t)$ where σ is the conductivity associated with the remaining normal electrons.

In summary of this complete Londons' formulation, we have

$$J_S(x,t) = J_{SE}(x,t)+J_{SN}(x,t_t)+J_{SB}(x,t_t)$$

where

$$J_{SE} = \Lambda^{-1}\int_{t_t}^{t} E(x,t)\,dt,\ J_{SN} = fJ_N(x,t_t-),\ \text{and}$$

J_{SB} is defined by (a) $\nabla\times J_{SB} = -\Lambda^{-1}B(x,t_t)$, (b) $\Delta\cdot J_{SB} = 0$, (c) the normal component of J_{SB} on the surface of a superconductor is zero and (d) the surface integral of the normal component of J_{SB} over a simple cross-section is zero.

These relations have been obtained on the basis that Λ^{-1} is uniform throughout the superconductor and unchanging in time. If two superconductors S_1 and S_2 abut, wherein Λ_1^{-1} is greater than Λ_2^{-1}, it is reasonable to assume that the current distribution is the superposition of one distribution, corresponding to a superconductor of shape that of a combination of S_1 and S_2 with characteristic coefficient equal to Λ_2^{-1}, upon another distribution corresponding to a superconductor of shape S_1 and coefficient $(\Lambda_1^{-1}-\Lambda_2^{-1})$. Then the preceding boundary conditions stated for J_{SB} apply successively to the two superposed distributions. Thus at the interface, the normal components of J_{SB} in S_1 and J_{SB} in S_2 are equal and the tangential components can be discontinuous. However, it does not seem possible to extend the foregoing considerations to instances of continuous spatial dependence of Λ without introducing postulates in addition to those already given. Thus the equation $\nabla\times J_{SB}(x,t_t) = -\Lambda^{-1}(x)B(x,t_t)$ cannot be valid for arbitrary forms of $\Lambda^{-1}(x)$, since the right-hand side, like the left, must be divergenceless. The Londons wrote equations of this kind as $\nabla\times(\Lambda J) = -B$, but there appears to be no satisfactory physical justification for placing the Λ inside the differentiation operators. It will be necessary to touch upon this subject again later, but considerations based on a constant Λ will suffice for now.

C. Questions of causal mechanisms

The formulation of the Londons' equations, stated in the preceding section as they were obtained, primarily from the literature but with rather appreciable alterations and refinements, constitutes a complete set of induction relations.

The currents induced in a superconductor of constant Λ, are defined uniquely, although perhaps inaccurately from an experimental point of view. The initial conditions and the boundary conditions play an essential part in achieving this required uniqueness. Furthermore, the relations were so constructed as to show the operation of two separate causal mechanisms, the one being effective only during the transition. Of course, this does not mean that the same fundamental laws do not give rise to both mechanisms, and naturally it is expected that the quantum electrodynamics of solids contains ultimately an explanation for all of them.

With the application of the Faraday–Lenz field relation

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

and a differentiation of the foregoing induction relations (neglecting the terms J_{SN} and J_N for the present), one can immediately see that

$$\nabla \times J_S(x, t) = -\Lambda^{-1} B(x, t)$$

In the literature it is this equation and the one already postulated

$$\frac{\partial J_S(x, t)}{\partial t} = \Lambda^{-1} E(x, t)$$

which are most often given as the pair of Londons' equations constituting the phenomenological induction relations for a superconductor. In view of the discussion of Section IIIA, however, it is seen that this is not an optimum form because it contains a field relation too, and is hence not a direct statement of pure induction relations. If this point were overlooked, these equations would superficially appear as describing two separate phenomenological mechanisms operating continuously and at all points in the superconductor. The agent for one of these mechanisms would be the magnetic field and the agent for the other would be the electric field. Such a conclusion, of course, does not bear up under examination because of the field relation between E and B on the one hand and the fact that the same dependent variable J_S appears in both equations.

We have here an essentially linear theory. The electromagnetic field obeys in every significant respect the principle of superposition, and the Londons' equations are also linear as to all magnitudes. Thus, if one cause, say a given magnetic field, produces a certain effect, a certain current distribution, when acting alone, and if a second cause alone, say a given electric field, produces another effect, another current distribution, then we should expect that when both causes act together the effect would be the linear superposition of the two separate effects, that is a current distribution which is the sum of the other two. The present form of the Londons' formulation refutes that expectation. In considerable measure, the joint effect is no more than either of the two separate effects taken alone.

Since the Maxwell field relations establish a close connection between the electric field and changes in the magnetic field, an induced superconductive current related directly to the time integral of the electric field is correspondingly related to all changes in the magnetic field subsequent to the transition to the superconducting state, or vice versa. But at the time of the transition no current can flow as a result of time integrals, which are then necessarily zero, and so the magnetic field must be presumed to produce a current distribution by itself. Then, in a causal sense, we might suppose the current is always produced by the magnetic field only, and the electric field relation is merely a convenient reformulation. Unfortunately, this cannot be done because an electric field may be produced apart from a magnetic field change, and then it produces an additional superconductive current. Thus, we are forced to say that the magnetic field produces a current at the instant of transition, and ever after, new magnetic fields produce nothing while the changes in current are produced by the electric field, or to say that the magnetic field always produces current but only that electric field not caused by magnetic field change produces additional current. The former of these modes of description was already given a complete formulation in the preceding section. The latter mode is susceptible to a rather similar formulation, which is given in identical notation below. However, it is quite difficult to conceive of a physical mechanism which, in any small region of the material, would respond differently to an electric field depending on whether that field owed its origin to electro-

static charges or to magnetic induction. Hence this formulation is very probably not useful for inferring what basic mechanisms are involved. The total superconducting current density

$$J_S(x, t) = J_{SB}(x, t) + J_{SE}(x, t) + J_{SN}(x, t_t)$$

where the first two terms on the right are defined as follows:

(a) $$\nabla \times J_{SB}(x, t) = -\Lambda^{-1} B(x, t),$$

(b) $$\nabla \cdot J_{SB}(x, t) = 0,$$

(c) where n is the normal to a surface and x' lies on that surface, $J_{SB}(x', t) \cdot n = 0$ if that surface is the surface of the superconductor, and

(d) $$\int_S J_{SB}(x', t) \cdot n \, dS = 0$$

if that surface is a simply connected cross-section. (Surface functions are here always interior.)

(a) $$\nabla \times J_{SE}(x, t) = 0,$$

(b) $$\nabla \cdot J_{SE}(x, t) = \Lambda^{-1} \int_{t_t}^{t} \nabla \cdot E(x, t) \, dt,$$

(c) where n and x' are as above,

$$J_{SE}(x', t) \cdot n = \Lambda^{-1} \int_{t_t}^{t} E(x', t) \cdot n \, dt$$

if that surface is the surface of the superconductor, and

(d) $$\int_S J_{SE}(x', t) \cdot n \, ds =$$

$$= \Lambda^{-1} \int_S \int_{t_t}^{t} E(x', t) \cdot n \, dt \, dS$$

if that surface S is a simply connected cross-section. J_{SN} was defined in the previous section.

In the preceding paragraphs, the existence of two phenomenologically distinct mechanisms was made evident. It is interesting that two mechanisms are still evident in the vector potential mode of formulation. Thus in

$$J_S(x, t) = -\Lambda^{-1} A(x, t) + J_{SN}(x, t_t)$$

apart from the J_{SN} term defined as before, there is one apparent mechanism in the induction of the current density J_S by the vector potential A. However, a given, physically unique electromagnetic-field configuration can be described with equal precision by any one of an infinite number of different vector potentials. The theoretician can choose any one of these in order to facilitate his calculations. But the current induction mechanism in the superconductor responds to only a particular one. There is, therefore, another mechanism, or a distinct, additional part of the first, which performs this selection process. It is indeed difficult to imagine any physical process which could discriminate between physically equivalent modes of description, and thus we reinforce our earlier conclusion that, despite its conciseness, the vector potential is not best suited for formulating induction relations. The correct vector potential may be calculated from the field and whatever boundary conditions are appropriate to a superconductor, and in every respect this amounts to solving the same equations and using the same boundary conditions as in the preceding formulation.

The formulations of the induction relations given here and in the preceding section are all quantitatively equivalent. They show that from a phenomenological point of view there are irreducibly two distinct mechanisms characteristic of a superconductor, neither one being immediately derivable from the other. The formulations given in this section, however, are decidedly inferior to that of the preceding section in providing insight as to just what those mechanisms might be.

In conclusion, we identify neither perfect conductivity nor perfect diamagnetism as more fundamental than the other, but we find very strong physical reasons for expecting a more fundamental understanding under the assumption that perfect diamagnetism is active only during the transition process, and only perfect conductivity prevails throughout the life of the superconducting state.

It is also quite conceivable, and existing experimental data do not yet appear adequate to exclude the possibility altogether, that both of the aforementioned types of superconduction mechanisms are active at all times. In this event the Londons' equations

and
$$\nabla \times J_S(x\,t) = -\Lambda^{-1}B(x, t)$$
$$\frac{\partial J_S(x, t)}{\partial t} = \Lambda^{-1}E(x, t)$$

to use the more conventional form and to neglect the terms J_{SN} and J_N, or the PIPPARD modification, would be quite inaccurate in certain kinds of experimental situations. The corresponding phenomenological equations would instead have to be formulated as follows:

$$J_S(x, t) = J_{SE}(x, t)+J_{SB}(x, t) \quad \text{(neglecting } J_{SN}\text{)}$$

where

$$J_{SE} = \Lambda_E^{-1}\int_{\dot{t}t}^{t} E(x, t)\,dt$$

as before, but J_{SB} is no longer an integration constant. It is defined by:
(a) $\nabla \times J_{SB} = -\Lambda_B^{-1}B(x,t)$, now time dependent, while the other conditions are as before:
(b) $\nabla \cdot J_{SB} = 0$, (c) the normal component on the surface is zero, and (d) the surface integral of the normal component over a simply connected cross-section is zero.
Possibly Λ_E^{-1} need not be identical with Λ_B^{-1}, and if in some materials the latter were nearly zero, one would have the description of superconductors with practically no Meissner effect, even though the materials were perfectly homogeneous. It is to be noted that from a phenomenological point of view, unequal Λ's could similarly occur in the formulation of Section IIIB, with similar descriptive consequences implied.

D. Large-scale non-local aspects

PIPPARD[8], in introducing the concept of coherence, and in modifying the Londons' equations so that these induction relations expressed the induced current at a point in terms of a kind of spatial average of the field throughout the neighborhood, clearly produced a formulation which is non-local. The range of this non-local averaging is relatively large on an atomic scale, but still microscopic in terms of bulk superconductors. However, the Londons' equation itself, that is, the curl equation for J_{SB}, is also a non-local formulation. While the equation for J_{SE} relates the current density at a point to the history of the electric field at that same point, the equation for J_{SB} is, as BARDEEN[13] states, not a point relation, since the current density at a point depends on the magnetic field in the surrounding region. Indeed, the current at a point is determined in substantial part by boundary conditions prevailing in exceedingly remote regions. The non-locality, in this case, takes in the entire bulk of the superconductor.

The extent of this non-localness presents a significant obstacle to a fundamental understanding of the phenomena. Classically, the response of a given volume element of a medium can come about only through the electromagnetic field present and acting upon it. The volume element responds to what is going on elsewhere only when the corresponding field conditions finally reach it by propagation. Neglecting mechanical or acoustical forces, the one exception to this is the continuity relation between charge and current, for the conditions in a volume element do depend on what flows in and out, but the Londons' relation is not of that kind. Consequently, one might well conclude that it is the quantum-mechanical wave function, in effect another field variable, which is sensitive to conditions everywhere, and, so to speak, transmits the information to the volume element. Hence, with SCHAFROTH and BLATT[14], one would require a specimen-wide correlation distance, however implausible that might otherwise seem, in order to obtain a Meissner effect as described by the Londons' equations.

This peculiarity of the Londons' equations is even more striking when one applies the formulation to two blocks of identical superconducting material in contact over a single, simply connected area. It might be supposed that, if the formulation were applied to each of the two blocks separately, merely with care being taken that the boundary conditions at the interface matched, precisely the same results would be obtained as by applying the formulation to the two blocks taken together as a single aggregate. Surely the physical situation is identical in the two approaches. The results differ.

E. Speculation on a model for the Meissner effect and the transition process

In Sections IIIB and C it was shown that if the

Londons' phenomenological description of superconductors was correct, even apart from small-scale discrepancies, there were strong and perhaps compel ing physical reasons for ascribing the Meissner effect, that of magnetic-flux expulsion on transition into superconduction, to a mechanism which is effective during the transition process but completely inactive thereafter, although the superconducting currents generated by it will carry on indefinitely. Since, as a matter of general observation, the transition into the superconducting state, at least in the presence of a magnetic field, takes place by a very pronounced nucleation and growth process, it is conceivable that this growth, involving as it does the motion of a boundary, provides just the mechanism sought. Whether the particular mechanism postulated later in this section can in fact be justified quantitatively, and on a fundamental atomic basis, remains to be seen. At least it does satisfy this apparently very important condition of being active only during the transition.

In further consideration, a process of nucleation and actual growth of superconducting regions, if a fundamentally essential feature of the transition process, removes the apparent inconsistencies of the Londons' formulation in regard to interface boundary conditions, as mentioned in the preceding Sub-section D. Once two bodies are in contact (sufficient to obtain continuity of wave functions), this growth process can continue along and across the interface, with obviously different ultimate results than if two different nucleation and growth processes came up to the interface from either side but remained separated and distinct. In much the same way, multiple connectedness can be handled without special treatment and becomes devoid of fundamental significance.

Moreover, it is no longer necessary to suppose a specimen-wide correlation distance to see that a configuration such as the Londons', or better, the PIPPARD modification, will be approached as the limit of growth. Considering the superconducting state as perfectly passive, such that a current configuration once established will persist indefinitely unless altered by an electromagnetic field, one sees that the growth process permits a point by point readjustment of the current distribution to the bounding field conditions as they are encountered and changed during the enlargement of the superconducting region. There may or may not be short-range coherence of wave functions, but the long-range coherence is effected electromagnetically, much as the positions of the bricks in a brick wall are correlated. The correlation of remote regions, once achieved in this way, is preserved only because of the complete permanence of the configuration; thus in the brick wall the top and bottom courses are correlated only because of the rigidity of the intervening material. A subsequent change in the superconducting-current pattern (apart from quenching), once the transition is completed, is produced solely by the electric field, and we have seen there is no long-range order implied in the induction relation appropriate to this.

For purposes of constructing a specific model, let us assume that a superconducting region contains charge carriers, whether they are fermions, fermion pairs or bosons, which behave so as to constitute a non-viscous fluid such that currents set up in it, by the momentary application of some agency of force, such as an electric field, will continue to circulate in it unchanged in configuration, for an infinite period of time, while the fluid remains contained by the walls of the region. Thus, the region displays perfect conductivity.

We suppose that the material, presumably in a magnetic field, as it approaches the conditions in which the superconducting fluid will form, passes at first into the superconducting state only in the neighborhood of points where the conditions or characteristics of the material are slightly more favorable or receive favorable fluctuation effects. This would probably denote the appearance of an appreciable energy gap at these points. These regions now act as sources of the superconducting fluid which flows or diffuses (not necessarily a random-walk process) outwards into regions which are slightly less favorable to the formation of the fluid but where the fluid, having once formed, can still remain intact or only partially attenuated. Let us say that these regions would have a finite, but temporarily much smaller, energy gap. If the transition probability into superconduction increased with gap size, then the regions of large gap could more quickly generate a concentration of fluid than the regions of smaller gap, and hence become sources for the latter.

Now in this process of outward diffusion, the fluid passes across the lines of magnetic induction and correspondingly receives lateral thrusts which bend the fluid into current paths surrounding the central source regions and the regions immediately adjacent. These currents, as is easily seen, are in such directions as to weaken the magnetic fields on the insides, and hence the inside regions not formerly able to generate the superconducting fluid can now do so, or if previously able slowly, can now do so more rapidly, thereby contributing to the further outward diffusion and the further expulsion of flux. To compensate for the slight charge imbalance resulting from this outward diffusion, an inward diffusion of the normally conducting fluid will take place. However, the latter, being closely coupled to the lattice, will perhaps move more slowly and in any event will transfer the thrusts it receives from the magnetic field to the lattice, so that there will be negligible counter diamagnetic currents generated. At the same time that the outward diffusion into the magnetic field creates peripheral diamagnetic currents, the interior acquires the properties of a bulk perfect conductor. The diamagnetic currents persisting there from the previous growth stage are neutralized by the counter electric field produced by the new currents.

Inhomogeneities in material or temperature distribution, or changes in the latter with time, correspond to space and time variations of Λ in the Londons' formulation and require a generalization of the treatment which, though of simple mathematical form, appears without good physical reason. The present condensation and diffusion mechanism would appear potentially capable of a very natural treatment of such variations, because they would correspond to variations in condensation and diffusion rates.

The experimental results of MEISSNER[15] and the theory of PARMENTER[16], while not precisely demonstrating the diffusion phenomenon attributed here to the transient state, clearly show that the superconducting fluid can, in the steady state, intrude long distances into a material which taken by itself would not be superconducting. A simple experiment to verify the present model would be to construct a superconductor by evaporation upon a substrate cooled well below the critical temperature and in the presence of a less-than-critical magnetic field perpendicular to the substrate. Since here the superconducting regions can grow only along the field lines, because of the direction of deposition of new material, no Meissner-effect type of diamagnetism or flux expulsion should occur, nor yet a transition into the intermediate state.

F. Concluding summary

Under the assumption that superconductors conform to the Londons' phenomenology, excepting small-scale effects relating to surfaces and high-frequency properties, strong evidence shows that two distinctive mechanisms operate to produce the characteristic behaviour. One of these is effective only during the normal-to-superconducting transition, and is to be viewed as responsible for the Meissner effect. If the nucleation and growth process is fundamentally associated with this mechanism, a number of deep-lying difficulties, such as multiple connectedness, extremely long-range coherence, and choice of gauge, tend to disappear. Furthermore, spatial and temporal inhomogeneities of material or temperature distribution probably require no special postulates or other unnatural disturbances of the straightforward application of the physical concepts involved.

It is conjectured that the process of condensation and diffusion of the perfectly conducting fluid may be the specific mechanism active during the transition process. This idea therefore indicates a program of further theoretical investigation to elucidate transition probabilities and rates of condensation and to ascertain the existence and magnitudes of the supposed diffusion or flow process. In the event of an outcome along the lines conjectured, it would be expected to obtain a fairly complete explanation of the Meissner effect, a good quantitative description of the electrodynamic transition process, and even a better understanding of static superconductivity as well, with the possibility of relaxing some of the strained portions of present fundamental theory which have had to deal with the difficulties mentioned above.

REFERENCES

1. A. B. PIPPARD, *Phil. Mag.* **41**, 243 (1950).

2. T. E. FABER, *Proc. Roy. Soc.* A**219,** 75 (1953).
3. W. B. ITTNER, *Phys. Rev.* **111,** 1483 (1953).
4. J. I. GITTLEMAN and W. H. CHERRY, *Bull. Amer. Phys. Soc. II* **3,** 383 (1958).
5. J. M. BREMER and V. L. NEWHOUSE, *Phys. Rev. Letters* **1,** 282 (1958).
6. R. F. BROOM and E. H. RHODERICK, *Brit. J. Appl. Phys.* **11,** 292 (1960).
7. F. LONDON, *Superfluids* Vol. I. Wiley, New York (1950).
8. A. B. PIPPARD, *Proc. Roy. Soc.* A**216,** 547 (1953).
9. J. BARDEEN, L. N. COOPER and J. R. SCHRIEFFER, *Phys. Rev.* **108,** 1175 (1957).
10. F. G. WERNER and D. R. BRILL, *Phys. Rev. Letters* **4,** 344 (1960).
11. D. SCHOENBERG, *Superconductivity* p. 180. University Press, Cambridge (1952).
12. J. A. STRATTON, *Electromagnetic Theory* p. 250. McGraw-Hill, New York (1941).
13. J. BARDEEN, *Handbuch der Physik* XV (Edited by S. FLÜGGE) pp. 299 and 319. Springer, Berlin (1956).
14. M. R. SCHAFROTH and J. BLATT, *Phys. Rev.* **100,** 1221 (1955).
15. H. W. MEISSNER, *Phys. Rev. Letters* **2,** 458 (1959).
16. R. H. PARMENTER, *Phys. Rev.* **118,** 1173 (1960).
17. W. A. LITTLE, *Canad. J. Phys.* **37,** 334 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 306–311. Printed in Great Britain

BRITISH RESEARCH ON THE CROWE CELL

D. H. PARKINSON

Royal Radar Establishment, Malvern, Worcestershire

INTRODUCTION

WORK on superconductive switching devices which could be used in computers has been going on in Britain since 1957. Three laboratories have been concerned, R.R.E. and S.E.R.L., who have been working on Crowe cells, and N.P.L., whose interest is in strip cryotrons. More recently industrial laboratories such as Mullards have become involved. The work at N.P.L. is reported after this paper, and S.E.R.L. are reporting on a device which has developed from their work; in consequence this paper is concerned with the work on CROWE cells[1]. At R.R.E. our objective has been to develop a storage device such that the random access time to a store containing 10^7 bits or more would not be limited by the switching speed of the device itself but by the time taken for a pulse to traverse the store once only. A store of this size composed of Crowe cells could probably occupy a 35-cm cube and the transit time of a pulse would be 1 mμsec. Allowing for the selection system, a switching speed of about 5–10 mμsec is a reasonable aim.

Early investigations of their properties showed that Crowe cells made using tin deposited on mica could be switched considerably faster than those made from lead or from tin on other substrates such as glass; drive pulse lengths of about 10 mμsec were adequate for reversing the cross-bar current. Tin is also a convenient metal because the critical current can be adjusted by varying the helium-bath temperature ($T_c = 3{\cdot}7$°K). With drive currents up to about 500 mA, pick-up signals of between 10 and 50 mV were observed. Investigations have focussed on the problem of how a cross-bar is driven normal and on those factors which govern the switching time. Attention is now turning to questions of reproducibility and to investigations of ancillary devices for gating and switching.

SWITCHING OF CROWE CELLS

The construction of the cells is relevant to a discussion of switching times. Those at R.R.E. have been prepared by depositing the film with circular holes, changing masks and then putting down the cross-bars (600 Å–1000 Å thick). A thin coating of an insulating material (silicon monoxide or magnesium fluoride) has been used as a protective layer. Pick-up (or drive) strips have been evaporated on the reverse side of the mica, the return path being through the main film itself. Drive (or pick-up) strips have been evaporated on a second sheet of mica, protected by an insulating layer, aligned accurately and held in contact with the front surface of the cells. At S.E.R.L. silicon monoxide insulating layers have been evaporated on the front surface of the cells followed by the drive strips.

In a discussion of switching speeds three time constants must be considered. Firstly, the electrical time constant, L/R, which governs the rate of current decay when the cross-bar becomes resistive. For the 2-mm diameter tin cells with cross-bars 0·2 mm wide and 2000 Å thick the self inductance is typically 10^{-9}, while R, which may be only a fraction of the full normal resistance of the cross-bar, is about 0·1 Ω. Therefore L/R is about 10 mμsec.

The other two time constants are thermal, and govern the rate at which the heat, produced in the cross-bar during switching, can be removed, first to the substrate (the fast time constant) and then to the helium bath (the slow time constant). This heat flow from the specimen has been investigated[2] by preparing cross-bars or evaporated strips with current and potential leads attached. The following measurements were then made: (i) resistance against temperature using a very small direct current (e.g. 200 μA), (ii) resistance against current at a constant bath temperature

less than critical and (iii) resistance against time during the application of short, square current pulses at the same bath temperature. Typical results are shown in Fig. 1. In measurements such as (ii), thermal instability occurs except when working very near the transition temperature.

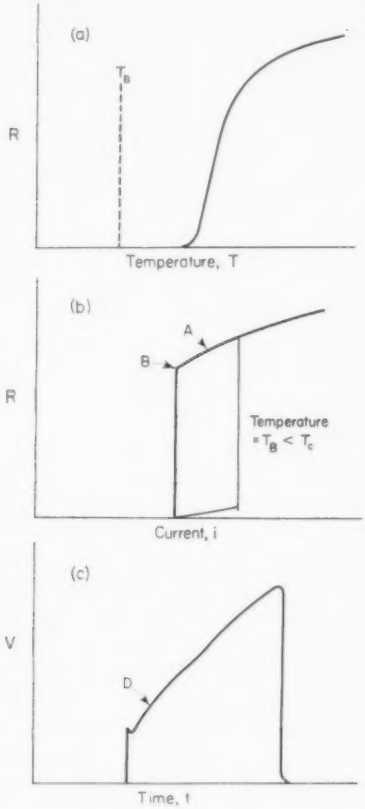


FIG. 1.

For the analysis of the heat flow a simplified model has been used which is shown in Fig. 2, with the symbols used in the equations which follow. The resistivity and temperature are assumed to be uniform throughout the tin film. Then for a point such as A in Fig. 1(b) the heat generated is just balanced by the heat conducted to the helium bath and

$$i^2R = (T-T_b)/(W+W_s) \qquad (1)$$

but T cannot be greater than the temperature corresponding to the same resistance point A on the low-current curve, Fig. 1(a). Therefore an upper limit can be placed on $T-T_b$ and hence on $(W+Ws)$. Also at point B, where instability is imminent, with the model of Fig. 2,

$$(i^2\,dR/dT)(W+W_s) \simeq 1 \qquad (2)$$

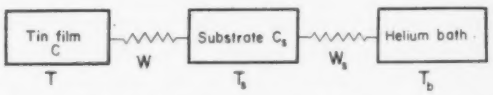


FIG. 2. (C,C_s = thermal capacity of tin film and substrate, respectively; W, W_s = corresponding thermal resistances; T, T_s, T_b = temperature of tin, substrate and helium-bath, respectively.)

It is unlikely that (dR/dT) at a resistance R and current i is greater than that at resistance R and smaller current. If dR/dT is taken from curve 1(a) it should be a maximum value and so give a lower limit for $W+W_s$. Such an analysis leads to an estimate of about 1 W/cm² per °K for the thermal conductance between the tin strip and helium bath.

In the pulse measurements, using a very fast rising edge of ~2–3 mμsec, the resistance of the strip rises very rapidly in about the first 20 mμsec and then much more slowly towards the full normal resistance [Fig. 1(c)]. The initial rapid rise may be attributed to the heating of the tin strip alone, and the slower rise to the heating of the substrate. From the numerical values it can be deduced that the thermal resistance W between tin and substrate is less than that W_s between substrate and helium bath. During the early part of the slow rise [e.g. point D in Fig. 1 (c)] well before equilibrium is approached, it may be assumed approximately that

$$i^2R = (C+C_s)\,dT/dt = (C+C_s)(dT/dR)(dR/dt)$$

Using a reasonable value for dT/dR, an estimate of $C+C_s$ can be made and is at least 1000 times greater than the calculated heat capacity of the tin alone. This supports the idea that the slow rise is due to the heating of the substrate, and leads to an estimated slow time constant $(C+C_s)W_s$ of tens of microseconds. An upper estimate of the fast rise time, which on this model is characterized by the time constant CW, can be obtained from the calculated value of C and the value of

$W_s + W$ ($> W$) derived from the d.c. measurements. This gives $CW < 30$ mμsec which is in substantial agreement with the observations.

Thus the fast thermal and electrical time constants are roughly equal, and together they determine the minimum length of drive pulse which can be used for satisfactory switching. Experimentally, cells with cross-bars about 1000 Å thick can be switched with 100–300 mA pulses of about 10 mμsec duration. The fastest times reported are from S.E.R.L. where cells with cross-bars of 600 Å thickness have been switched in about 3 mμsec.[3]

The analysis also suggests that the slower switching of lead films may be due to the higher heat capacity of the lead and a lower conductance to the substrate caused by poorer acoustic matching.[4]

Theoretical calculations based on a simplified model[5] (e.g. neglecting the latent heat involved in the superconducting to normal transition) have also been used to investigate the behaviour of a cell. The self inductance L, thermal resistance to the substrate W and heat capacity C were taken as constants. The cross-bar resistance R was assumed to be zero for currents less than I_c, and for currents greater than I_c it was assumed to be the following function of temperature and current:

$$R/R_n = 1 - \exp[-f\{(I/I_c)-1\}]$$

where $f \sim 0{\cdot}6$ and

$$I_c/I_0 = 1-(T/T_c^3)$$

I_c and T_c are the critical current and temperature, and I_0 the critical current at absolute zero. The differential equations for the cross-bar current I and temperature T have been integrated numerically on the T.R.E.A.C. electronic computer. The behaviour of the cross-bar depends on these quantities in a complicated way, but the calculations show that the flat top of the drive pulse must be longer than the L/R or fast thermal time constant CW to avoid undesirable effects such as signals arising from the trailing edge of the pulse. Fig. 3 shows the results of a calculation. The variation in critical current in regions such as PQ, which is due to heating when the cross-bar goes normal, depends on the short thermal time constant. The fall in the cross-bar current RS depends on the L/R time constant. The residual stored current (when the pulse is removed) is dependent on the pulse shape. Conditions can be found in which the stored current does not reach a reproducible value until 10–20 pulses

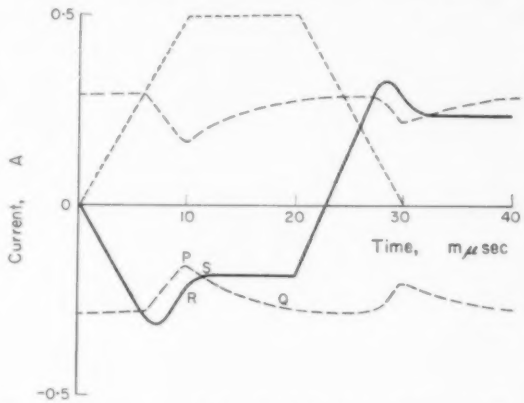


FIG. 3. Computed typical pulse behaviour of Crowe cell (- - - - - drive current; ——— cross-bar current; — — — critical current).

have passed. Although calculations such as these are over simplified, they serve as a very useful guide to the detailed behaviour of the cells.

The residual current in a cross-bar has also been investigated experimentally[6] by cycling a cell with first a positive and then a negative drive pulse as shown in Fig. 4. Between them a short pulse can be injected and increased in amplitude until the pick-up signal is disturbed. The amplitude of the positive injected pulse which is just sufficient to disturb the system measures $I_c + I_r$ where I_r is the residual stored current in the cross-bar, while a negative pulse measures $I_c - I_r$; hence I_r can be found. The value to which the cross-bar current falls during a pulse is given by $I_d - I_r = I_a$, where I_d is the drive current. These relations are obvious in Fig. 4. As expected from the calculations, the values of I_a and I_r depend on the shape and amplitude of the drive pulse. At this stage it is difficult to be certain just how much of a disadvantage this is likely to be.

In the analyses given above it has been assumed that in the restoration of resistance the restored part is distributed over the whole cross-bar. At the other extreme the restoration of resistance could be due to the existence of one or more defects or weak points[7] at which normal resistance first

appears and from which the heat generated spreads. The truth probably lies between these two extremes: there may be many weak points in a cross-bar. To test this idea a strip, with potential

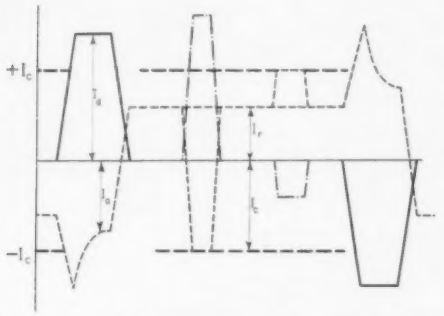


FIG. 4. ——— Drive current; - - - - cross-bar current; ·—·—· injected pulse.

leads attached centrally and at either end, was examined and with steadily increasing pulses the return of resistance in each half was compared. So far, results show that there is little difference between the two halves, consistent with the existence of a large number of weak points uniformly distributed.

EVAPORATION TECHNIQUES AND REPRODUCIBILITY

Cells have been made in a number of different ways, ranging from melting and evaporating tin from directly heated boats to pre-melting in vacuum and then dropping the molten pellet into a boat heated by electron bombardment. Deposition rates ranging from 100 Å to 1000 Å per sec have been used according to the technique of evaporation. Those films which have been evaporated faster have the best appearance to the eye but tend to have transitions which are spread out (as wide as $\sim 0{\cdot}2°K$ between the 90 and 10 per cent resistance values). These are objectionable, amongst other things making the pulse sizes more critical for coincident current working. The more slowly evaporated cells have a dull appearance to the eye but narrower transitions.

Most of the work referred to has been carried out on individual cells made in batches of 3×1 or 3×3 in one set of evaporations. In the earlier experiments remasking to lay down the cross-bars was carried out by breaking the vacuum. More recently, the complete cells have been produced without breaking the vacuum. Evaporating conditions and masking techniques are being improved rapidly and studies of the factors governing the reproducibility of the cells are going on at the Mullard Research Laboratories. At present, not nearly enough is known about the proper conditions for preparing satisfactory films.

THE USE OF CROWE CELLS IN COMPUTERS

As far as speed is concerned the switching time of the faster Crowe cells is compatible with pulse lengths of about 100 mμsec, which was the original object. To use such cells in a computer a persistent current must be stored for long periods without deterioration. In Fig. 5 the results of a typical test are shown in which the current has been satisfactorily stored for a 10-min interval. It is also shown there that the current is virtually unaffected by many half read-or-write pulses and hence the cells can be used in coincident current working.

In preliminary tests individual cells have been switched continuously and satisfactorily with 10 Mc/s sine waves.[2] In a cell with an inductance of $\sim 10^{-9}$ H and critical current ~ 100 mA, repeated switching at 20 Mc/s would dissipate $\sim 2 \times 10^{-4}$ W. If on an average as many as 50 bits are switched at once (100-bit words) the total power dissipation in the store would be $\sim 10^{-2}$ W, which is acceptable. What is more serious is the local rise in temperature at a cell which is repeatedly switched at the maximum frequency of say 20 Mc/s. Allowing for the cross-bar to substrate area $\sim 0{\cdot}004$ cm^2, there would be a dissipation of about 50×10^{-3} W/cm^2. In practice this figure may be higher, say 0·1 W/cm^2 to allow for higher switching speeds and smaller areas of the cross-bar going normal; on the other hand sideways spread of heat in the film and the substrate could reduce it. From the thermal conductance to the helium bath 0·1 W/cm^2 yields a temperature rise of $\sim 0{\cdot}1°$K, which is probably the most which could be tolerated.

To be able to use Crowe cells satisfactorily in a store, gating devices, selection systems and registers have to be devised. There are a number of devices based on semiconductors which are fast

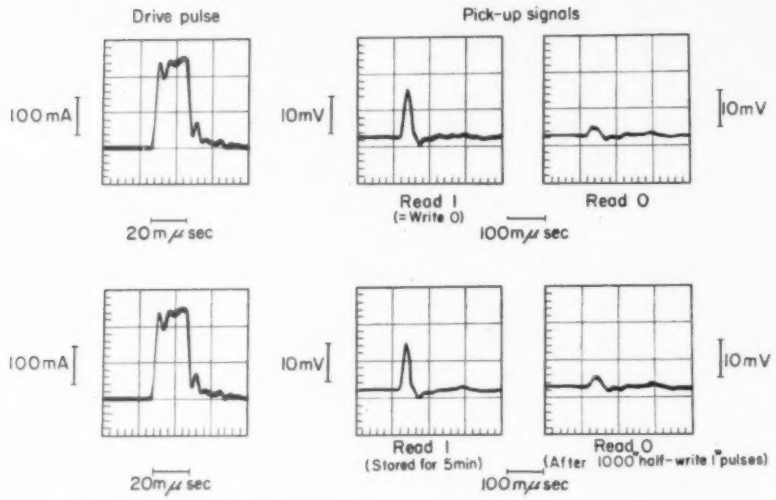


FIG. 5. Response of 2-mm tin Crowe cell at 2·6°K. (Drawings of pick-up signal photographs are of "single-shot" scope traces.)

enough to be suitable for use under liquid helium. An elementary selection system[5], based on the cryosar, has been tested, and p–n-junction diodes work well at helium temperatures, the elbow of their characteristic sharpening so that a rise of only 0·3 mV will increase the current by a factor of e in the region of the elbow. The signal from the Crowe cell ($\sim$ 20 mV) has also been used to switch a suitably loaded tunnel diode.[5] It is feasible in this way to build a register to handle the rather weak signals from the store. It is now possible to propose on paper a store in which much of the selection and gating is carried out using devices based on semiconductors.

In Fig. 6, A and B are two germanium cryosars on which similar arrays of contacts are placed. On to one face of A a set of conducting strips has been evaporated, each strip corresponding to a row of words. On the other face behind each strip there are a set of contacts leading to the drive wires, one to each word. All the wires in one column of words are then connected together at the far end. By selecting the xth wire on the front of A and the yth at the far end of the store, a pulse can be driven down the xyth word. Similarly, any digit in the word can be selected using plate B from which the wires are connected to the digit-pulse generator. The Crowe cell has been shown to be suitable for coincident current working, so the combination of plates A and B give the basis for writing in the store. Interrogation can be carried out with plate A by selecting a word and driving a full amplitude pulse along it. The pick-up array is shown in Fig. 7. The digit gates could be p–n-junction diodes, probably made of indium antimonide, and the digit register composed of tunnel diodes. This scheme is obviously very much in its early stages of conception and there are several matching and transmission-line problems to be solved.

Finally, some effort has been directed to the design of a single-stage expansion-engine helium refrigerator based on experience gained with a two-stage machine similar to the A.D. Little Collins Liquefier.

The philosophy behind the store described here is that the whole computer including the arithmetical elements shall be at helium temperatures and, once cooled, the computer should as far as possible be kept cold. "Trouble shooting" under these conditions is not an insuperable problem; with proper design, remote-handling techniques could enable units to be tested whilst remaining under helium, and also enable units to be removed from a helium bath and replaced. However, small prototype stores have yet to be built.

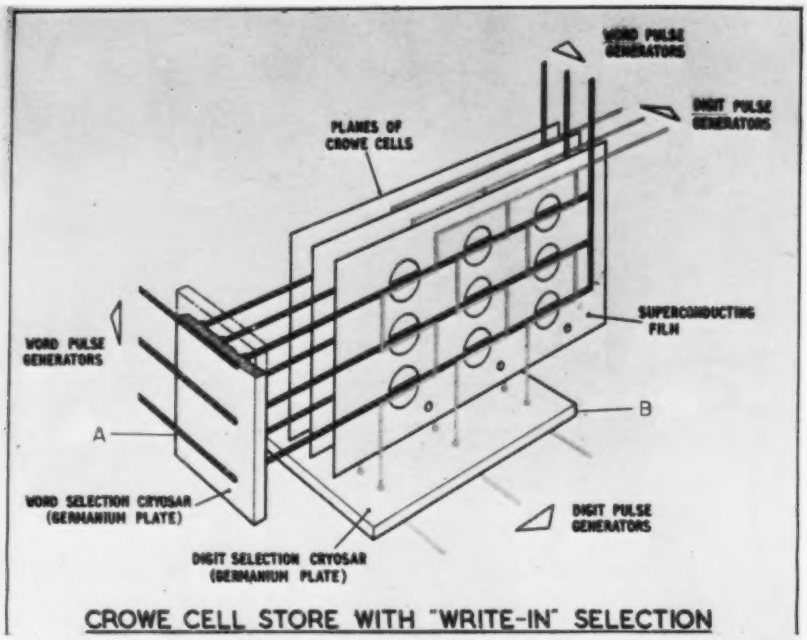


FIG. 6.

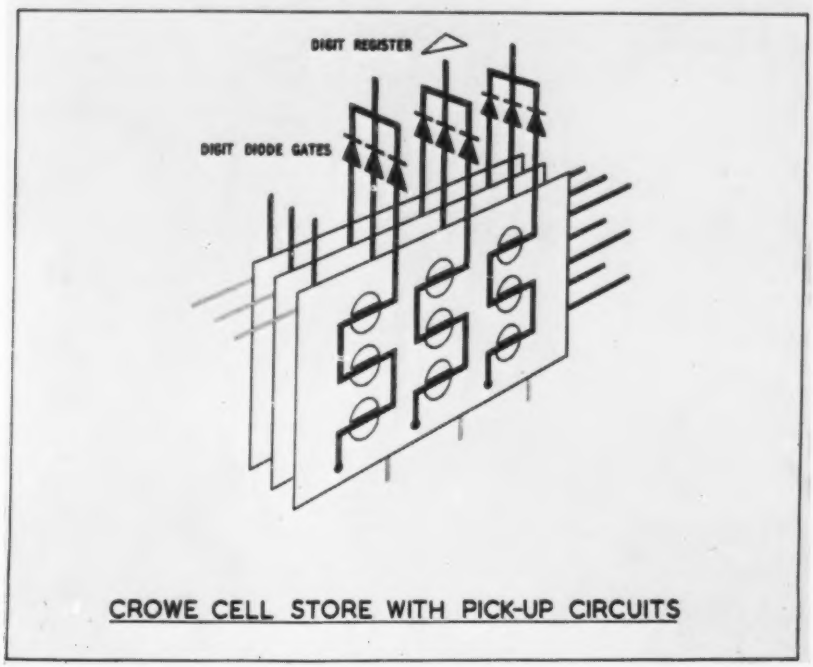


FIG. 7.

FIG. 6. Crowe cell store with "write-in" selection.

FIG. 7. Crowe cell store with pick-up circuits.

REFERENCES

1. J. W. Crowe, *I.B.M. J.* **1,** 195 (1957).
2. L. M. Roberts, Research at R.R.E. To be published.
3. E. H. Rhoderick and R. F. Broom. Private communication.
4. W. A. Little, *Canad. J. Phys.* **37,** 334 (1959).
5. J. M. Lock. Private communication.
6. E. H. Rhoderick and R. F. Broom. Private communication.
7. R. F. Broom and E. H. Rhoderick, *Solid-State Electron.* **1,** 314 (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 312–313. Printed in Great Britain.

BRITISH RESEARCH ON THIN-FILM CRYOTRONS

P. R. STUART

National Physical Laboratory, Teddington, Middlesex

SUPERCONDUCTIVE computing elements have been studied at the National Physical Laboratory since October, 1958, with particular emphasis given to the planar cryotron. It was soon apparent that one of the greatest problems would be that of reproducibility. Accordingly, work has been concentrated on a study of the factors determining the transition temperature and the sharpness of the superconductive transition of thin films of tin.

Usually the tin films have been deposited in pairs of 5 mm × 0·1 mm strips, each strip having separate current and potential leads.

It has been found that if two strips are simultaneously deposited on a sheet of glass which has been cleaned with methyl alcohol and by ion bombardment, their transition temperatures may differ by as much as 0·01°K. If, however, a layer of silicon monoxide is deposited first, the transition temperatures are usually within 0·002°K of one another. When an 800 Å film of silicon monoxide was deposited under only one strip of the pair, the transition temperature of this strip was 0·045°K below that of the strip deposited straight on to the glass. When both strips were deposited on a silicon monoxide layer and one was overlaid with a further film of silicon monoxide, its critical temperature was raised 0·01°K above that of the unprotected layer.

These effects are probably due to strain caused by differential contraction on cooling, and these results show that careful control of the deposition of the insulating films of a device is just as important as the control of the deposition of the superconductors.

Specimens annealed at 150°C, or deposited on a substrate at about 150°C, have had a residual resistance ratio of about 0·01, compared with the value of 0·025 usually obtained for specimens deposited on a substrate at room temperature. The superconductive transition range (0·1 R_n to 0·9 R_n) was, however, increased by both these means to 0·035°K, a sixfold to tenfold increase on the values obtained for specimens deposited at room temperature. A specimen deposited on a substrate at about 90°K and allowed to warm up to room temperature on transferring from the evaporater to the cryostat had a transition temperature of 3·83°K, or 0·07°K above the average for specimens deposited at room temperature. Its residual resistance ratio and transition spread were much the same as for room-temperature deposition.

Our evaporation chamber contains a liquid-nitrogen-cooled vapour trap, ensuring that the pressure during evaporation is less than 10^{-5} mm mercury. A film deposited at the relatively high pressure of 5×10^{-4} mm mercury had a residual resistivity of $1{\cdot}7 \times 10^{-6}$ Ω-cm, compared with values in the range $0{\cdot}15$–$0{\cdot}45 \times 10^{-6}$ Ω-cm obtained in the trapped system. The transition temperature of this film was, however, within the range normally encountered.

It has been noticed that the transition temperature of a tin film changes with the aging of the film. For example, the transition temperature of an unprotected tin film aged in the atmosphere for 13 days increased by 0·02°K, while that of a simultaneously deposited strip aged in helium at room temperature increased only 0·007°K.

Our thin-film cryotrons have been made in a vacuum evaporation unit which has four vapour sources, on a rotatable turret, and a mask holder which will hold up to twelve masks. Resistance-free contact between the films can only be made if the vacuum remains unbroken between the deposition of successive films.

The cryotrons, which have been based on glass substrates have had, in general, a tin gate of width 2·4 mm and thickness 3000–7000 Å and a lead control strip of width 0·1 mm. Lead backing films

and fired silver-paste contacts have been used. Silicon monoxide and magnesium fluoride have both been found unsatisfactory as insulators, the former often wrinkling when deposited over a tin film, and the latter cracking when deposited over a lead film. Silicon monoxide is more satisfactory if evaporated slowly and at a relatively high pressure of oxygen.

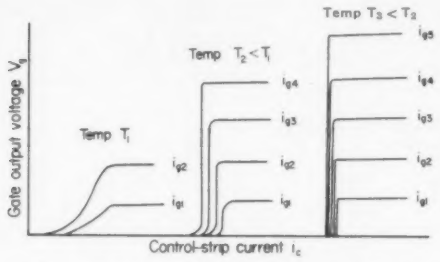


FIG. 1. Cryotron switching characteristics.

Six cryotrons have recently been tested, and their switching characteristics, obtained by the use of a transistor chopper amplifier coupled to an X–Y plotting table, have the form shown in Fig. 1. In these characteristics, V_g denotes the voltage across the gate when it is carrying a current i_g, and i_c denotes the current through the control strip.

Acknowledgement—This work was carried out as part of the research programme of the National Physical Laboratory and is published by permission of the Director of the Laboratory.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 314–318. Printed in Great Britain

A NEW TYPE OF BISTABLE ELEMENT INVOLVING THERMAL PROPAGATION OF A NORMAL REGION IN A THIN SUPERCONDUCTING FILM

R. F. BROOM and E. H. RHODERICK

Services Electronics Research Laboratory, Baldock, Hertfordshire

1. INTRODUCTION

THE work described in this paper is a continuation of earlier experiments with thin superconducting films of lead and tin, driven into the normal state by rectangular current pulses.[1] An explanation of the results of these experiments was advanced in which the growth of the normal region was attributed to a thermal spreading mechanism of the sort described by NEWHOUSE and BREMER[2]. The velocities of propagation observed by them were less than 100 cm/sec, whereas the velocities required to explain our results are of the order of 10^6 cm/sec. It was suggested[1] that this large difference might be due to the much higher current densities used by us. The present work originated as an attempt to analyze the thermal spreading process in greater detail, to find out whether it leads to propagation rates of the right order of magnitude to explain our experimental results.

2. THE THERMAL SPREADING MECHANISM

The problem which we wish to solve is that of the effect of a current passing through a thin superconducting strip, an arbitrary length of which is in the normal phase. If the joule heating of the normal region is so great that the adjacent superconducting material is raised to the critical temperature, the interphase boundary will move, and it is this boundary movement which we wish to investigate. The problem is a difficult one which is probably insoluble without the use of a computer, and we shall therefore consider a simpler problem which can be easily solved using standard mathematical techniques, but which is sufficiently close to the real problem to give some insight into the thermal spreading mechanism.

We consider the problem of an infinitely long strip carrying a current I, half of which is in the normal and half in the superconducting state, and inquire how the interphase boundary moves as a result of the joule heating of the normal region. Suppose that the resistance of the strip in the normal phase is r per unit length, and that the strip loses heat to its surroundings at a rate $\alpha w\theta$ per unit length, where w is the width of the strip and θ the difference between the temperature of the strip and the bath temperature T_0. (We assume that the strip is sufficiently thin for the temperature variation normal to its surface to be negligible.) Measurements by BROOM and SIMPSON[3] of thermal relaxation times of lead films on various substrates imply that the loss of heat from the film is predominantly to the substrate, even when the free surface of the film is in direct contact with liquid helium. The constant α is therefore probably determined by the thermal resistance between film and substrate due to the acoustic mismatch at the interface.[4] In this paper we shall deal with times of the order of a microsecond or less, during which period there is insufficient energy available to cause significant heating of the substrate. The latter will therefore be regarded as an infinite heat sink at the same temperature as the bath.

At large distances from the interphase boundary in the normal region, the excess temperature θ approaches the value $\theta_1 = I^2r/\alpha w$ at which the joule heating just balances the loss of heat to the surroundings; at large distances from the boundary in the superconducting region, the temperature of the film approaches that of the substrate so that $\theta = 0$. The temperature distribution along the

strip at any instant of time therefore resembles that shown in Fig. 1, the interphase boundary occurring at x_1, where $\theta = \theta_c$. θ_c is thus the excess temperature above the bath temperature at which the strip becomes normal when carrying a current

With these assumptions it is easily shown[5] that a solution of the required form exists, and that the velocity of the interphase boundary is given by

$$v = \{(\alpha K/d)^{1/2}/c\rho\}(\theta_1 - 2\theta_c)\{\theta_c(\theta_1 - \theta_c)\}^{-1/2}$$

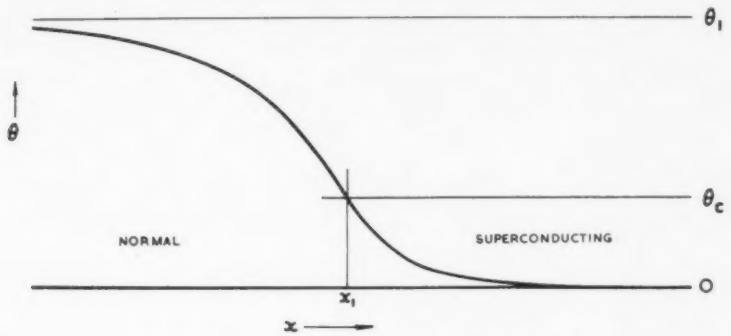


FIG. 1. Variation of temperature along the length of an infinite current-carrying strip, part of which is superconducting and part normal.

I, and is a decreasing function of I which vanishes when I becomes equal to the critical current corresponding to the bath temperature.

If we assume that the conduction of heat along the length of the strip can be represented by the usual thermal conductivity equation, which is not really justifiable when the thickness of the strip is less than the phonon mean free path, the differential equations governing the temperature distribution along the strip in the normal and superconducting regions can be easily written down.[5] To solve them, we assume that there is a solution having the character of a travelling wave, i.e. one in which the interphase boundary moves with constant velocity v and the temperature profile remains unchanged relative to the boundary, so that $\theta = \theta(x - vt)$. In seeking this solution we make the following simplifying assumptions: (1) It is assumed that, as the temperature distribution changes with time, the boundary moves so that it always occurs where $\theta = \theta_c$, and that there is no time lag in this process, i.e. the possibility of any relaxation effects in the phase transition is neglected. (2) The existence of a latent heat of transition is neglected. (3) The specific heats and thermal conductivities of the film in the superconducting and normal phases are assumed to be equal and independent of temperature.

where K is the thermal conductivity of the film, d its thickness, c its specific heat and ρ its density. The novel feature of this result is that v is positive or negative, i.e. the normal region grows or collapses, according as θ_1 is greater or less than $2\theta_c$. If $\theta_1 = 2\theta_c$ the boundary remains stationary, and it is easily confirmed that for this particular case a time-independent solution of the differential equations is possible. As I approaches the critical current I_c corresponding to the bath temperature, θ_c approaches zero and θ_1/θ_c tends to infinity. Since θ_1 and θ_c are both monotonic functions of I, it follows that the current I_1 for which $\theta_1/\theta_c = 2$ is less than I_c, so that the current required to maintain the boundary stationary is less than the critical current. The ratio I_1/I_c is determined by the numerical values of I_c, r, α and w, and by the amount the critical temperature (T_c) of the film exceeds the bath temperature T_0, but in all cases I_1/I_c is less than unity. Fig. 2 shows how the function $(\theta_1 - 2\theta_c)\{\theta_c(\theta_1 - \theta_c)\}^{-1/2}$ varies with I for the particular case of $I_1/I_c = 0{\cdot}5$. In constructing this graph we have assumed a linear variation of critical current with temperature, which is approximately true for temperatures close to T_c, so that $\theta_c = (T_c - T_0)(I_c - I)/I_c$.

To find the order of magnitude of v, consider the case of the lead strips used in our previous

experiments.[1] For these $d \sim 500$ Å, and if we take the bulk values for the thermal conductivity and specific heat appropriate to the normal phase near 4°K we obtain $K \simeq 10$ W deg^{-1} cm^{-1} and

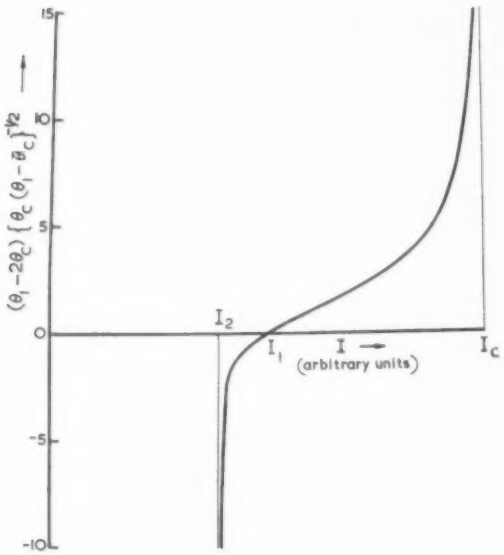


FIG. 2. Plot of $(\theta_1-2\theta_c)\{\theta_c(\theta_1-\theta_c)\}^{-1/2}$ against I, for the case $I_c = 2I_1$.

$c \simeq 8\times10^{-4}$ J deg^{-1} g^{-1}. Experiments in this laboratory with lead films on mica substrates have yielded values of 1–10 W deg^{-1} cm^{-2} for α, and 3 W deg^{-1} cm^{-2} will be used as a typical value. The factor $(\alpha K/d)^{1/2}/c\rho$ now becomes $\sim 3\times10^5$ cm sec^{-1}. Fig. 2 shows that $(\theta_1-2\theta_c)\{\theta_c(\theta_1-\theta_c)\}^{-1/2}$ tends to $+\infty$ as I approaches I_c, and to $-\infty$ as I approaches the value I_2 for which $\theta_1 = \theta_c$. Hence, in principle, any velocity of the boundary between the limits $\pm\infty$ is obtainable. In practice, since no film is perfect and θ_c must vary slightly along the length of the strip, one would not expect the theory to hold when I is close to I_2 or I_c. The fact that $(\alpha K/d)^{1/2}/c\rho \sim 3\times10^5$ cm sec^{-1} means that propagation velocities of the order of 10^6 cm sec^{-1} are obtainable without taking I very close to I_c, and the thermal spreading mechanism therefore seems capable of giving an adequate explanation of our previous experiments. In view of the simplifying assumptions made in the analysis and the uncertainties in the values of the constants involved, one cannot expect to be able to make an exact quantitative check of the theory.

The current I_c is the "true" critical current of the strip at temperature T_0, i.e. it is the current which would drive the film into the normal state if it were a perfect film and there were no heating effects. In our previous paper[1], we explained our results by assuming that, as the current is increased, the strip first goes normal at a flaw of some sort, which has a lower critical current than the rest of the strip. This flaw then heats up until, at a current I_c', it is hot enough to initiate the thermal propagation process and the normal region spreads throughout the entire strip. The current I_c' is the "apparent d.c. critical current", and, in the case of the lead films used in Ref. 1, I_c' may be as little as one-half of I_c. I' is not the same as the quantity I_1 introduced above. I' refers to a situation in which the length of the normal region is extremely small ($\sim 10^{-4}$ cm), while the result that the normal region grows for currents greater than I_1 was derived on the assumption that the normal region is semi-infinite in extent. (In practice this probably means that the normal region must be greater than $\sim 10^{-2}$ cm.) The difference between I_c' and I_1 is discussed more fully in Ref. 5, where it is shown that, for the films quoted previously[1], $I_c' > I_1$.* Thus, in Fig. 2, I_c' occurs somewhere between I_1 and I_c, and the conclusion that the current required to maintain the boundary stationary is less than the critical current remains true for our imperfect films. For a typical lead film 2 mm × 0·125 mm × 500 Å deposited on a mica substrate, $I_1 = 320$ mA and $I_c' = 550$ mA at 4·2°K, while I_c was estimated to be $\simeq 800$ mA.

3. APPLICATION TO A BISTABLE ELEMENT

The fact that I_1 is less than both the "true" critical current of a perfect film and the "apparent" critical current of our imperfect films enables a bistable element to be constructed, consisting of a superconducting strip carrying a continuous standing current I_1. Application of a pulse of current in the same direction as I_1, and of sufficient amplitude, would initiate a normal region which would spread to an extent determined by the

* The quantities I_c' and I_1 are denoted by I_1 and I_2 respectively in Ref. 5.

amplitude and duration of the pulse. At the conclusion of this "ignition" pulse, the normal region would be held constant by the standing current I_1. Application of an "extinction" pulse, equal to I_1 but in the opposite direction to the standing current, and of duration comparable with the cooling-time constant of the strip, would cancel the standing current for a time long enough for the strip to become completely superconducting once more, and the strip would remain in this state at the conclusion of the extinction pulse.

Bistable elements of this sort have been constructed using both lead and tin strips. There are two advantages to be gained by using tin rather than lead for such an element. Firstly, by working near the critical temperature the critical current may be reduced very considerably compared with that of a lead strip, which has the twofold advantage of reducing the amplitudes of the triggering pulses and of decreasing the power dissipation in the resistive state. Secondly, because of the smaller cooling-time constant of tin,[6] much shorter extinction pulses may be used. A typical tin strip with dimensions 2 mm × 0·125 mm × 450 Å, on a mica substrate, had $I_1 = 11$ mA and $I_c' = 28$ mA at 2·95°K. This was operated with a standing current of +11 mA, and ignition and extinction pulses of +22 mA and −11 mA respectively. Both pulses could be as short as 10 mμsec in duration. The amount of resistance generated by an ignition pulse of this amplitude was 0·9 Ω, while the full normal resistance of the strip was 6·6 Ω. The current and voltage waveforms for this strip are shown in Fig. 3, the repetition rate being 2 kc/s. The baseline corresponding to zero current has been drawn to show the standing current.

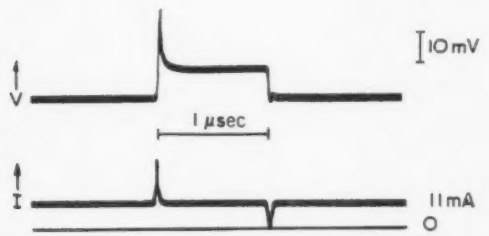


FIG. 3. Voltage and current oscillograms for tin strip 2 mm × 0·125 mm × 450 Å at 2·95°K. The ignition and extinction pulses are both of 20 mμsec duration. The zero-current baseline has been drawn to show the standing current.

The value of the standing current was not as critical as one might expect from the above theory, and a variation of nearly ± 10 per cent could be tolerated; as Fig. 2 shows, the dependence of velocity on current is relatively small near I_1. It was possible to operate this strip with the order of the pulses reversed, i.e. with the strip in the resistive state for 500 μsec and completely superconducting for 1 μsec. The joule heat generated during the "on" period did not raise the substrate temperature sufficiently to cause any appreciable change in the operating conditions. This may not be true if several such elements are in close proximity, but it should be possible to minimize such variations in substrate temperature by operation below the lambda point.

This bistable element could conceivably find some application in computer design. Among its advantages are extreme simplicity of construction, high speed of operation, low drive current and non-destructive read-out in the form of a d.c. voltage. Its obvious disadvantage is that it consumes power in the "on" state, which limits the number of such elements that can be used without excessive helium consumption to $\sim 10^4$. The name "calotron" seems appropriate to the device because of the important part played by heating in its mode of operation. The calotron is really multistable rather than bistable, since the amount of resistance restored to the strip is continuously variable, depending on the amplitude and duration of the ignition pulse. Fig. 4 shows waveforms

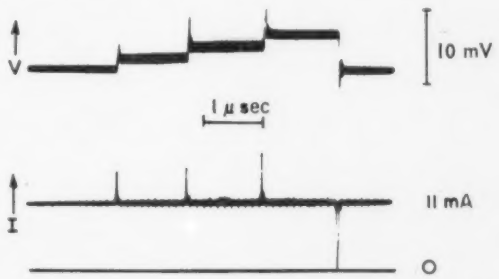


FIG. 4. Voltage and current oscillograms for tin strip 1 mm × 0·125 mm × 500 Å, with three ignition pulses.

obtained with a tin strip operated with three ignition pulses, the first generating a normal region which is subsequently enlarged by the other two

in turn. After each ignition pulse, the normal region is held constant at its new value by the standing current. In this mode of operation the device could be used as a digital-to-analogue converter.

4. CONCLUSION

The most important result of this paper is that thermal propagation of the interphase boundary in thin superconducting films carrying a current may be a much faster process than has previously been realized. As far as we are aware, this is the fastest known process in solids in which changes in lattice temperature play an essential role. Apart from its application to the design of new superconducting devices, it could well play an important part in the operation of existing ones. For example, in the planar cryotron the amount of resistance introduced into the gate by a current in the control strip just above the threshold value is usually $\sim 10^{-3}\ \Omega$ or less. If the gate strip carries a current close to its d.c. critical current, this small amount of resistance may be rapidly multiplied as a result of the thermal spreading process. In the Crowe cell, also, thermal propagation almost certainly makes an important contribution to the growth of resistance in the crossbar, while the reverse process in which the normal region collapses may play a part in its subsequent recovery. Since thermal relaxation times may be as short as 10 mμsec or less if an appropriate choice of film and substrate material is made, the use of a phenomenon in which heating plays an essential part is unobjectionable in many applications.

Acknowledgement—We are indebted to the Admiralty for permission to publish this paper.

REFERENCES

1. R. F. Broom and E. H. Rhoderick, *Phys. Rev.* **116,** 344 (1959).
2. J. W. Bremer and V. L. Newhouse, *Phys. Rev. Letters* **1,** 282 (1958).
3. R. F. Broom and O. Simpson, *Brit. J. Appl. Phys.* **11,** 78 (1960).
4. W. A. Little, *Canad. J. Phys.* **37,** 334 (1959).
5. R. F. Broom and E. H. Rhoderick, *Brit. J. Appl. Phys.* **11,** 292 (1960).
6. D. H. Parkinson, *Solid-State Electron.* **1,** 306 (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 319–322. Printed in Great Britain

INITIAL STUDIES OF THE PREPARATION AND PROPERTIES OF VANADIUM THIN FILMS

J. D. BLADES, J. GERBER and C. T. THOMPSON

Burroughs Corporation, Research Center, Paoli, Pennsylvania

INTRODUCTION

IT IS well known, from the work of WEXLER and CORAK,[1] that impurities located at interstitials have a major effect upon the ideal superconducting properties of vanadium. They have shown that as oxygen and nitrogen content in bulk samples increases, the superconducting transition becomes diffuse. Other important physical characteristics of vanadium, also, are sensitive to minute quantities of these impurities.[2] It is of interest to determine in a more direct manner the extent of this interstitial effect.

There are physically conceptual and experimental advantages to the consideration of this impurity problem in vanadium by utilizing the solid thin-film state. Transmission electron microscopy and resistance measurements are particularly amenable to this geometric case.

This paper reports some preliminary results of experimental studies of the preparation and the physical properties of vanadium thin films.

EXPERIMENTATION

Preparation of films

Thin films of vanadium, 300–3240 Å in thickness have been obtained by the vacuum-evaporation technique, utilizing a hot tungsten boat. Substrates, which were maintained at a desired temperature, were chemically cleaned glass, fused quartz and sapphire. The upper limit of the vacuum during deposition ranged from 6×10^{-7} to 10^{-4} mm Hg, after an adequate bakeout procedure, and was maintained by a conventional oil diffusion pump vacuum system. Pressures were measured by a Stokes Bayard–Albert ionization gauge. Evaporation times ranged from 2·5 sec to 19 min, and the boat temperature was monitored with an optical pyrometer.

Film surfaces are reflective and lack the vivid coloring usually characteristic of vanadium oxides. Film adherence was unsatisfactory only for the samples on the sapphire substrates. For electrical measurements, the film-sample geometry is a rectangle of 0·8 mm width and 1·35 cm length. Larger-sized samples were evaporated simultaneously and were utilized for structural and impurity investigations.

Physical measurements

The resistance of the vanadium thin films has been measured by a K-3 Leeds and Northrup potentiometer utilizing the four-prong connection procedure, at room, liquid-nitrogen and in the liquid-helium temperature regions. Temperature is monitored by Allen–Bradley resistance thermometers and by vapor pressures. A conventional closed dewar system with pressure regulation is utilized for the low-temperature measurements. A pulsing system has been designed and constructed, but experimental results so far are insignificant. Thickness measurements have been made by a Zeiss interference microscope.

RESULTS

The results that have been obtained are conveniently discussed in two sections. A structural investigation has been concerned with film texture and the electrical measurements with the residual resistance α, the ratio of the film resistance at liquid-helium temperature to the resistance at room temperature.

Film texture

According to HOLLAND[3] high melting point metals in the thin-film state, which includes vanadium, can be expected to have a micro-

crystal texture that is randomly oriented. VERDERBER[4] has verified this interpretation for nickel–iron thin films by the Geiger counter X-ray reflection technique; however, when the vanadium thin films were investigated, utilizing this same procedure, no reflections were observed. The reason for this is attributed to two important factors: (a) the intensity of the strongest vanadium reflection is 1/20th that of the strongest reflection from the nickel–iron lattice, as determined from standard materials, and (b) the possibility of an amorphous structure in the vanadium films at the substrate temperatures existing during deposition. Even if structure were present in the vanadium films, the relative intensity situation apparently dictates that it cannot be expected to obtain reflections by this X-ray technique, unless the film samples are thicker.

For the purpose of improving this experimental situation, we turned our attention to the electron microscope, and the results to be presented are representative of only two of the films that have been deposited. Fig. 1 demonstrates the surface condition of an 800 Å thick vanadium film. The micrograph is a chromium-shadowed collodion replica. It is obvious that no definite particle size or shape is apparent, and when compared to the oxygen-doped thin tin films of CASWELL[5], it should be anticipated that the vanadium film contains large amounts of impurities. This interpretation is further substantiated by the resistance measurements that have been made.

Fig. 2 presents a transmission electron micrograph of a 700 Å thick vanadium film with a vapor-beam angle of incidence of $\sim 30°$. The texture is similar to the iron and nickel–iron films of HEIDENREICH[6] except that a "line structure" is absent. However, upon closer examination one notices a slight, randomly directed "line or fiber structure". The film has been stripped from a glass substrate.

In Fig. 3, direct evidence of a micro-crystal texture is indicated by the transmission electron diffraction lines present. From the ring patterns, the *d* spacings of vanadium have been calculated, and this sample is the same film that was observed in Fig. 2. Further reorientation of this film with respect to the electron beam demonstrated no change in the rings; hence, the micro-crystals are randomly oriented.

Residual resistivity measurements

In Table 1 the pertinent data for fifteen evaporations of five vanadium films each have been collected. The initial pressure for each evaporation was in all cases less than the highest pressure

Table 1. *Data collected for fifteen vanadium thin-film evaporations*

Evaporation no.	Evaporation time (min)	Thickness (Å)	Substrate temp. (°C)	Highest pressure (mm Hg)	α
2	19	450	222	10^{-4}	1·02
3	6	2430	145	3×10^{-5}	0·77
4	11	540	150	4×10^{-5}	0·88
5	4	1350	105	5×10^{-6}	0·60
6	6	810	125	10^{-4}	1·05
8	19	2700	175	10^{-5}	0·94
9	3	1350	45	5×10^{-6}	0·64
11	0·75	3240	465	6×10^{-7}	0·65
*12	1·5	800	525	4×10^{-5}	1·26
13	2·0	540	250	5×10^{-5}	1·07
15	3·0	300	150	10^{-4}	1·10
16	0·5	2700	25	3×10^{-6}	0·65–0·69
17	7 sec	2430	20–25	1×10^{-5}	0·9
18	2·5 sec	540	95	1×10^{-5}	0·91
19	12 sec	2160	70	5×10^{-6}	0·70

*Oxide layer formed.

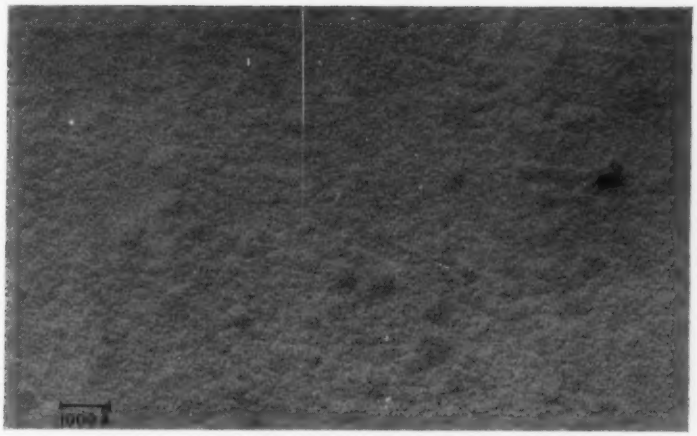


FIG. 1. A chromium-shadowed collodion replica electron micrograph of an 800 Å vanadium thin-film (120,000×).

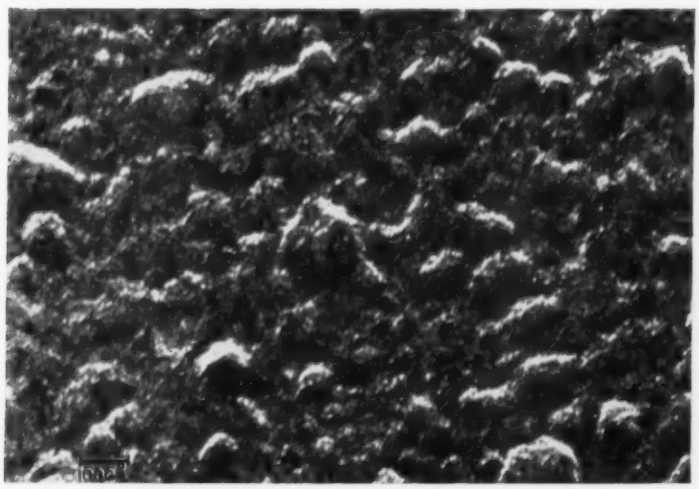


FIG. 2. A transmission electron micrograph of a 700 Å vanadium thin-film with incident vapor beam at ~30° (120,000×).

[*facing p.* 320

FIG. 3. A transmission electron diffraction pattern for the same vanadium thin-film as used in Fig. 2. The rings lead to the calculated *d* spacings for vanadium.

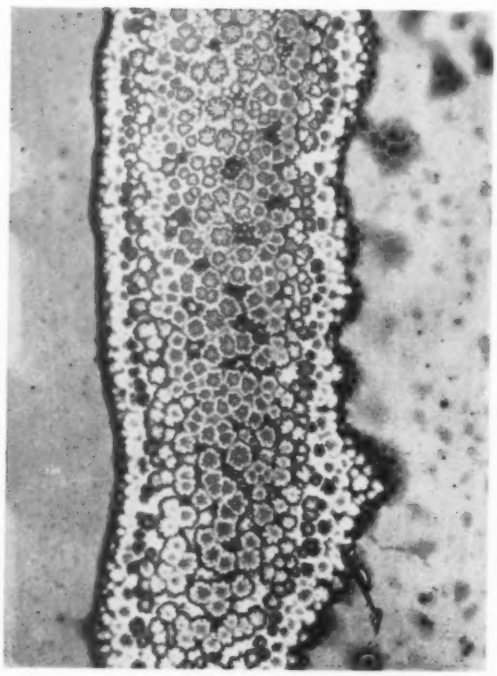


FIG. 4. An optical micrograph of apparent oxide-fiber structure in a vanadium thin-film with substrate at 525°C during condensation.

attained during the evaporation run. The residual resistance measurements are reproducible to within less than 1 per cent and the values in the table represent average values, in most cases, for all five samples, the deviation here being less than 1 per cent also.

high reflectivity and a low number of pit holes. Fig. 4 is an optical micrograph of the oxidized film and the traces of structural growth that occurred. Thickness variations account for the outlining of this structure, but no attempt will be made at this time to explain the structure.

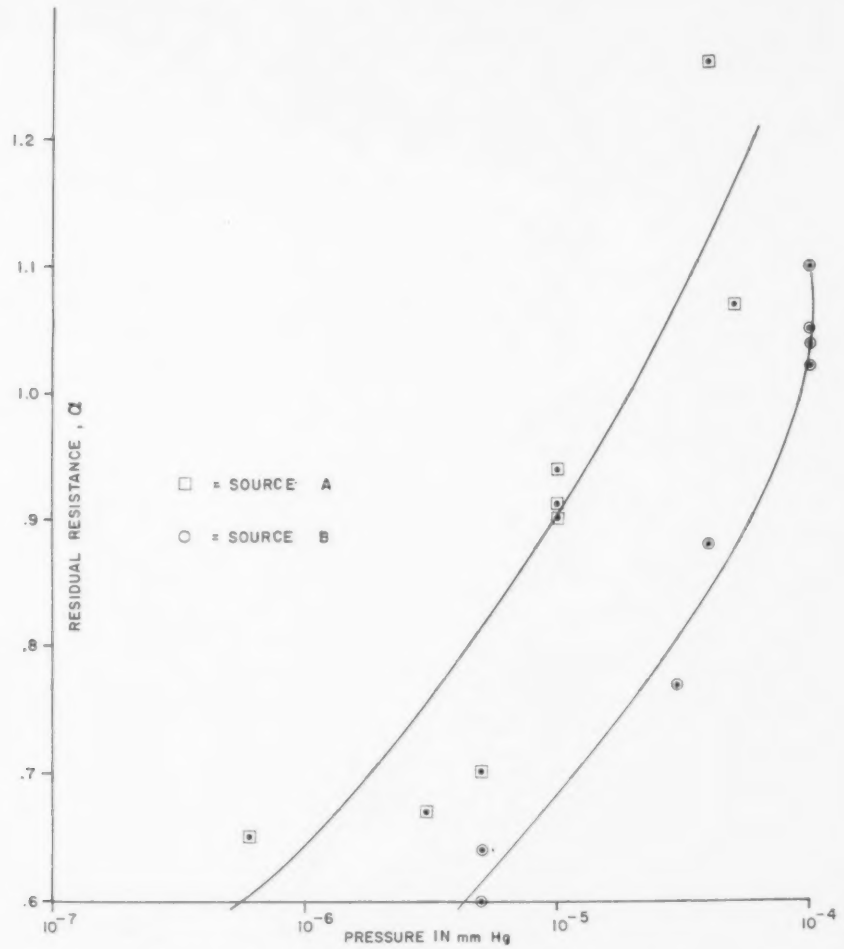


FIG. 5. Residual resistivity as a function of the highest pressure.

A number of features of the data are of importance and of interest. The hottest substrate sample formed an oxide layer, which was indicated by a golden color of its surface. All the other samples demonstrated reflectivity and, under a microscope, indicated good thin-film characteristics, such as

It is of further interest to note that, in some cases, the resistance of vanadium in the thin-film state at 4·2°K is larger than the resistance at room temperature, as indicated by α's>1. This is a situation that does not exist for the bulk state.[7] However, when the residual resistances are con-

sidered relative to the highest evaporation pressures, as shown in Fig. 5, it becomes apparent that the films are highly sensitive to vacuum impurity content.

The two experimental curves are for two distinct types of commercial bulk source materials. The pressure for $\alpha < 1$ is 3×10^{-5} mm of Hg and below. One point of the lower curve material is inconsistent, and this inconsistency is thought to result from the 19 min evaporation time. Resistance measurements upon samples of the bulk source material give $\alpha = 0{\cdot}09$, indicating that the superconducting state is attainable in these materials, and that probably the oxygen and nitrogen content is lower than in the films.

Evaporation 16 has a range in residual resistance of from 0·65 to 0·69. The five film samples were measured sequentially, beginning with the first one approximately 2 min after it and the other four films had been removed from the vacuum. As each sample was measured α increased. The last two samples measured almost identical values of α. This apparently is a weak indication of the effects of absorption by the film after being exposed to an atmospheric environment and because of the effects of time.

CONCLUSIONS

The experiments that have been discussed lead to the following conclusions in regard to the thin film of vanadium.

(a) The impurity content and the effects upon the superconducting properties at normal vacua are greater than in the case of bulk materials; however, there are indications that pure films might be obtained with an ultra-low vacuum and a special source preparation.

(b) A randomly oriented micro-crystal texture is found to exist in vacuum-deposited vanadium.

(c) No detectable relationship exists between substrate temperature and residual resistance.

(d) No detectable relationship exists between film thickness and residual resistance.

(e) High substrate temperatures may lead to heavily coated oxide films at the higher deposition vacua.

It should be re-emphasized that the results presented so far are of a preliminary nature. Further experiments are now in progress in an attempt to clarify the effects of interstitials upon the physical properties of thin-film vanadium.

Acknowledgements—It is our pleasure to acknowledge the welcomed and appreciated assistance of B. Rosenbaum, R. Alexander, R. R. Verderber and J. Cammarota. We are further indebted to C. Loog whose efforts made this work possible, and to the Office of Naval Research for our supply of helium.

REFERENCES

1. A. Wexler and W. S. Corak, *Phys. Rev.* **85,** 85 (1952).
2. G. A. Alers, *Bull. Amer. Phys. Soc.* **5,** 170 (1960).
3. L. Holland, *Vacuum Deposition of Thin Films* p. 207. Wiley, New York (1956).
4. R. R. Verderber, *J. Appl. Phys.* **30,** 1359 (1959).
5. H. L. Caswell, *Bull. Amer. Phys. Soc.* **5,** 164 (1960).
6. R. D. Heidenreich and F. W. Reynolds, *Structure and Properties of Thin Films* p. 402. Wiley, New York (1959).
7. J. A. Hren, C. M. Wayman and T. A. Read, *Some Properties of Vanadium at Subatmospheric Temperatures.* Air Force Office of Scientific Research Report TN59–1003 (1959).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 323–334. Printed in Great Britain

CURRENT-INDUCED SWITCHING OF SUPERCONDUCTIVE THIN FILMS*

F. W. SCHMIDLIN,†‡ ARTHUR J. LEARN,†
E. C. CRITTENDEN, JR.§ and J. N. COOPER§

CURRENTS required to cause superconducting-to-normal transitions of thin films have been measured as a function of temperature for various widths and thicknesses. Normal-to-superconducting transition times for the same films have also been measured as a function of the magnitude of drive current. In spite of the lack of a satisfactory theoretical model for the initiation of the transition, the bulk of the results can be understood in terms of temperature changes in the film, as governed by factors limiting the flow of heat (generated subsequent to the initial transition) into the substrate and liquid-helium bath.

TEST SPECIMENS

Specimens were prepared by evaporating indium or tin onto cold substrates of quartz or glass in a vacuum the order of 10^{-6} mm of Hg. Deposition rates were about 100 A/sec. A typical specimen is shown in Fig. 1. The substrates are optically polished quartz or glass disks 1 in. in diameter and 1/8 in. thick. The film edges were defined by razor blades. The material inside the irregularly shaped boundaries appearing in the wide contact areas at the ends of the film is indium solder, which was used for attaching a thermocouple and electrical leads. The latter enabled monitoring the conductance of the film during deposition. Film thicknesses were controlled by evaporating a known mass of metal to completion. Unfortunately, variations in boat shape led to gross errors and a better measure of thickness was obtained from the room-temperature resistance after correcting for contact areas and wall scattering (using SONDHEIMER's[1] theory).

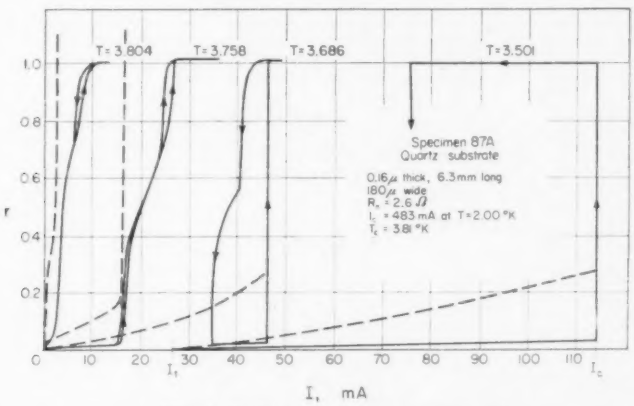


FIG. 2. Reduced resistance vs. current.

---

* Supported in part by the Office of Naval Research.

† Space Technology Laboratories, Inc., Los Angeles, California.

‡ Present address: Clauser Technology Corporation, Torrance, California.

§ U.S. Naval Postgraduate School, Monterey, California.

The thickness of a number of specimens was checked by multiple-beam interferometry which showed the resistance-determined thicknesses to be more reliable.

STEADY-STATE CHARACTERISTICS

The steady-state characteristics of a film were measured by passing a slowly increasing current through the film and observing the resulting voltage across the ends. For this purpose separate current and voltage contacts were provided. Results for a typical tin specimen are shown in Fig. 2. Here, the reduced resistance is plotted vs. current for four different bath temperatures. In each case the resistance begins to develop continuously, starting at the "threshold current" I_t, but terminates in a large abrupt rise at the so-called "d.c. critical current" I_c. The reduced resistance at I_c is the order of the width-to-length ratio of the film, 0·025 for specimen 87A. (Note the dashed curves in Fig. 2 with scale expanded 4 times.) At lower bath temperatures (below 3°K for specimen 87A), the reduced resistance at I_c was found to decrease with temperature, but at the λ-point of the helium bath it was suddenly restored to the order of the width-to-

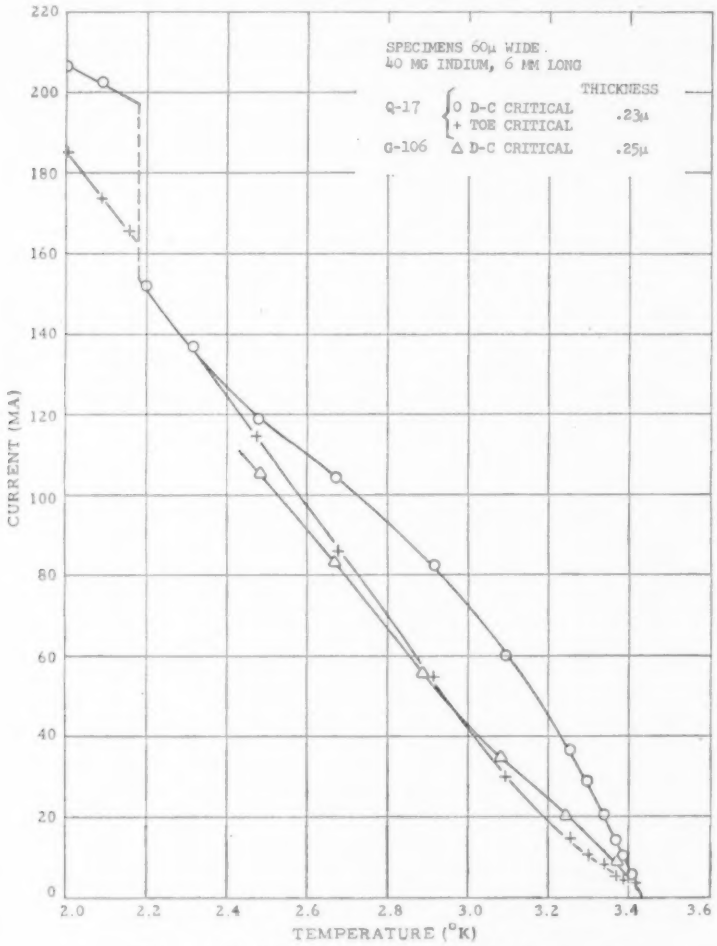


FIG. 3. Variation of critical current with temperature.

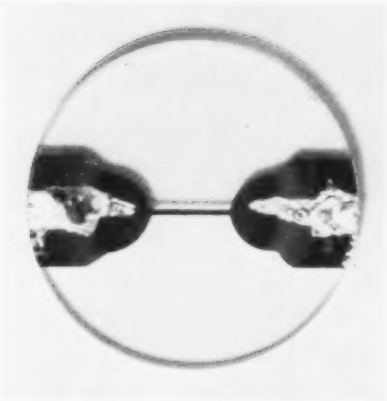


FIG. 1. Test specimen deposited on 1-in. diameter, optically polished glass disk.

[*facing p.* 324

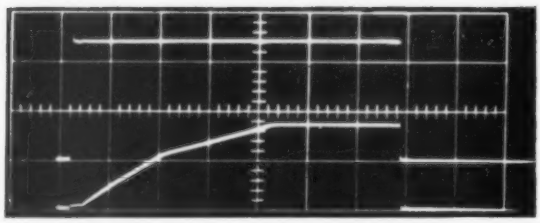


FIG. 8(a). Indium at 2·26°K. Voltage vs. time for switching at d.c. Critical current at temperatures above the helium λ-point. (Time scale: 100 μsec per large division.)

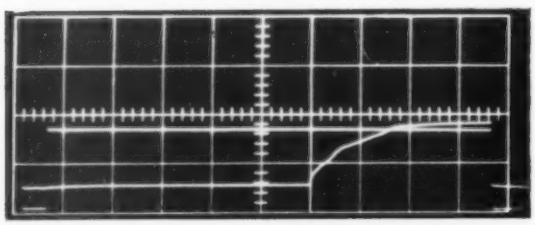


FIG. 8(b). Indium at 1·74°K. Voltage vs. time for switching at d.c. Critical current at temperatures below the helium λ-point. (Time scale: 2 μsec per large division.)

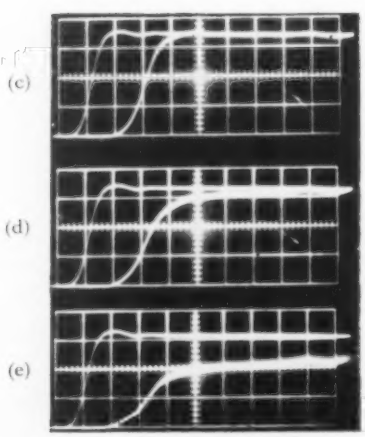


FIG. 8(c,d,e). Switching behavior for currents approaching I_s. (Time scale: 20 μsec per large division.)

length ratio. Indium specimens showed similar behaviour. The reduction of reduced resistance can be correlated with convergence of I_c toward I_t as the temperature is lowered. In Fig. 3, I_t and I_c are plotted vs. temperature for an indium specimen (Q-17) of geometry similar to that of assumed almost identical with that of Q-17. Another pair of specimens on quartz and glass substrates are compared in Fig. 4. No difference between I_c and I_t can be distinguished. In general, the continuous part of the resistance change for films on glass is negligibly small except very close

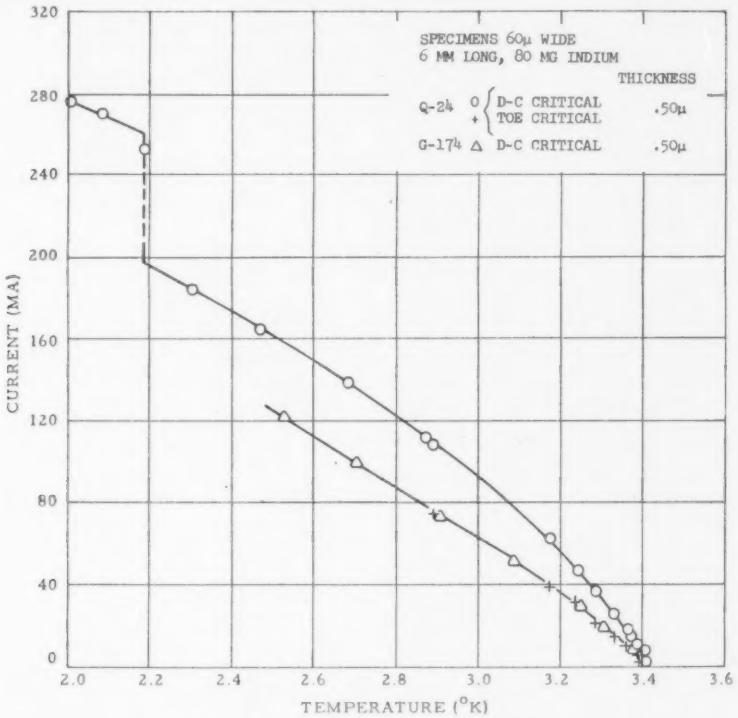


FIG. 4. Variation of critical current with temperature.

87A. The temperature at which I_c and I_t start to converge (about 3·1°K for Q-17) is of particular interest. It is dependent on film thickness as illustrated by comparing Figs. 3, 4 and 5—note that it decreases with increasing film thickness. For sufficiently thin specimens (e.g. Q-17 in Fig. 3), I_t and I_c actually coalesce. The region of coalescence corresponds to the discontinuous resistance change starting from a vanishingly small value. For specimens on glass, I_t and I_c coalesce much closer to the critical temperature. Critical-current curves for geometrically similar specimens on quartz (Q-17) and glass (G-106) are compared in Fig. 3. I_t for G-106, though not shown, can be to the critical temperature. The slight jump of I_t at the λ-point for Q-17 in Fig. 3 is an anomaly due to heat generated at the current contacts. In more careful experiments, where it was made certain that contact resistance was negligible (by adding a second pair of voltage leads which included the current contacts) no such anomaly was observed.

All of the above observations, as well as the increasing hysteresis with decreasing bath temperature shown in Fig. 2, can be understood on the basis of a heat-flow model. Fig. 6 shows a reasonable sketch of the thermal flux lines. For an order-of-magnitude calculation, the steady-state heat-

balance equation can be written in form

$$qA_s = h\Delta T\, A_{\text{eff}} \tag{1}$$

where $q = I_c^2\rho_n/W^2\delta$ is the heat generated per unit area, A_s = resistive area, h = heat-transfer coefficient at solid–liquid boundary, T = temperature increase, say at the center of the resistive area, A_{eff} = the effective area of the film and substrate through which heat flows, ρ_n = the normal resistivity of the film, W = film width and δ = film thickness.

To obtain the qualitative dependence of temperature rise on film geometry, the d.c. critical current as a function of width and thickness is needed. The measured results are shown in Fig. 7. Well within the experimental error, an expression

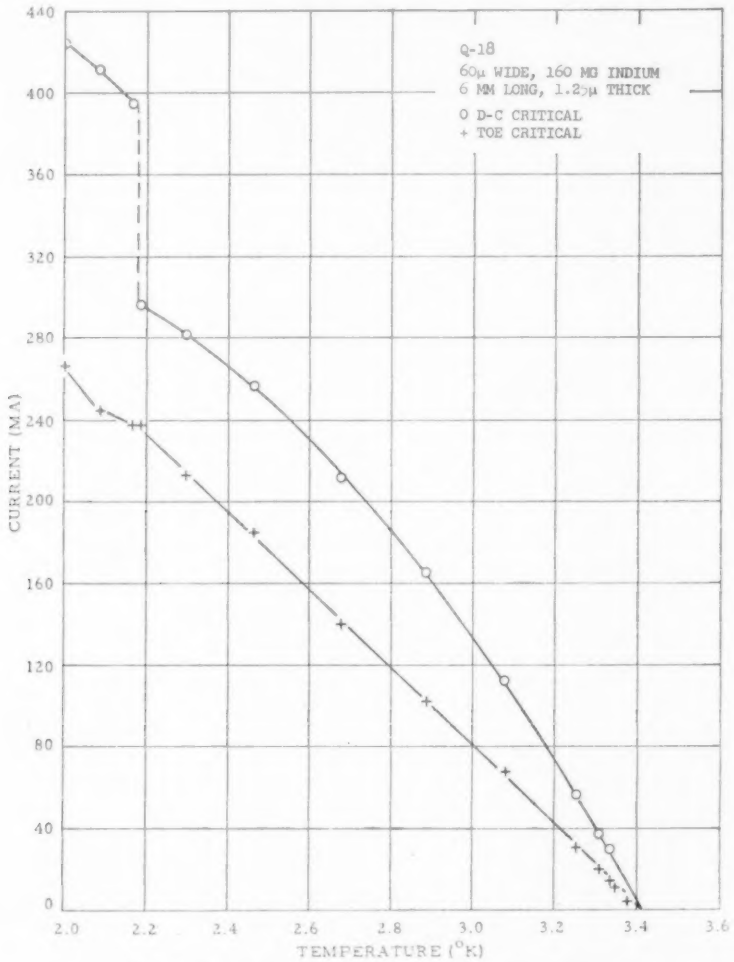


FIG. 5. Variation of critical current with temperature.

of the form

$$I_c = i_0\, W\sqrt{\delta} \tag{2}$$

fits the data. This relation was obtained for a bath temperature of 2°K to minimize thermal effects. Actually, as the present analysis shows, this particular geometrical dependence is partly due to joule heat. However, it is just the desired

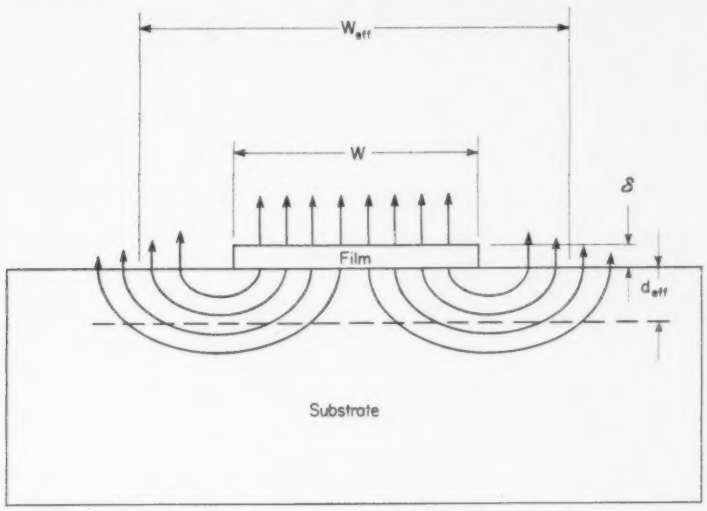


FIG. 6. Cross-sectional view of specimen showing thermal flux lines.

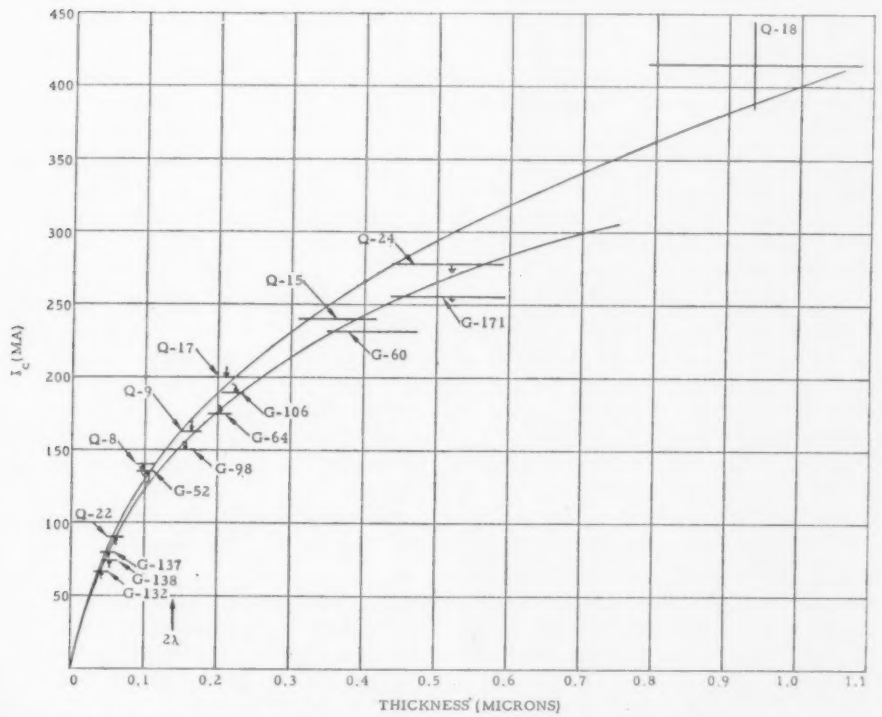


FIG. 7. D.C. critical current at 2°K vs. thickness of nominally 60-μ-wide specimens corrected to 62 μ. Upper curve for quartz substrates; lower curve for glass substrates.

expression for estimating the temperature increase at the point of instability.

Because of surface scattering, the dependence of film resistivity on thickness is, according to SONDHEIMER's theory[1],

$$\rho_n = \frac{\rho_0 l'}{\delta f(l'/\delta)} \tag{3}$$

ρ_0 is the bulk resistivity and l' is the mean free path of the electrons. f is a slowly varying function of thickness. By substituting equations (2) and (3) into equation (1), one obtains for the temperature increase in the normal region

$$\Delta T = \frac{1}{h}\frac{A_s}{A_{\text{eff}}}\frac{i_0^2 \rho_0 l'}{\delta f(l'/\delta)} \tag{4}$$

Now, if the instability in the transition occurs when ΔT reaches a critical value, then the observed features of the I_c vs. T curve can be understood as follows: The abrupt increase of resistance when the normal region grows to a length the order of the width of the film results from A_{eff} starting to grow in only one dimension instead of two. As the bath temperature is lowered, however, the power input increases and the heat-transfer coefficient decreases.[2] Hence the transition can become unstable even while A_{eff} is growing in two dimensions. In this region I_c converges toward I_t. At the λ-point of the helium bath, h suddenly increases so that conditions comparable to those near the critical temperature are restored. The inverse dependence on thickness δ explains the lower bath temperature required to achieve the above critical conditions in specimens Q-24 and Q-18 in Figs. 4 and 5 respectively. The large difference between the thermal conductivities of quartz and glass (a factor the order of 10^3), implies a much larger A_{eff} for quartz. Finally, it is of interest to obtain an order of magnitude of the temperature rise at the threshold of the instability. Assuming any superheating or supercooling effects negligible, A_{eff} can be estimated from the hysteresis. On glass A_{eff} was about $4A_s$, but on quartz it was about 200 A_s. Hence, for specimen Q-17 (Fig. 3) at 3°K and a current slightly less than I_c, ΔT is the order of 0·1°K, but under the same conditions on glass the temperature rise would have been 5·0°K.

Before turning to the transient characteristics, two more features of the current vs. temperature curves should be noted. First is the concave upward portion of I_t vs. T near T_c (Q-17 Fig. 3). This is due to the variation of the magnetic-field penetration depth λ with temperature. The magnitude of the effect is dependent on the thickness of the film relative to the penetration depth λ_0 at $T = 0$°K. For specimens Q-24 and Q-18 (Figs. 4 and 5), which are much thicker than twice λ_0, no effect was observable. The second is the portion of the I_c vs. T curve below the λ-point of the liquid helium. Here the curve can be fitted by an expression of the form

$$\frac{I_c}{I_0} = 1 - \left(\frac{T}{T_c}\right)^n \tag{5}$$

where n depends on thickness, ranging from 4 for very thin specimens ($\delta \ll 2\lambda_0$) to 2 for thick ones ($\delta > 2\lambda_0$). The same curve also fits the data near the critical temperature T_c.

TRANSIENT CHARACTERISTICS

By applying fast rise current pulses of amplitude greater than I_c, the abrupt resistance rise was resolved in time. Fig. 8 shows oscilloscope traces of voltage (and hence resistance) vs. time which were selected to illustrate three distinct types of switching behaviour. For reference, current pulses are also shown. Trace 8(a) for a current pulse of amplitude I_c and bath temperature above the λ-point, shows the resistance rise resolved into a time delay followed by an increase in linear segments. By increasing the gain of the preamplifier a small voltage can be observed during the time delay which corresponds to the continuous part of the resistance change. The latter occurs in less than 10^{-7} sec for pulse amplitude $I = I_c$. (It is less than 5×10^{-9} sec at $I = 1{\cdot}1\ I_c$, but very difficult to measure.) The linear segments correspond to normal regions sweeping along the length of the film as described by BREMER and NEWHOUSE[3] who observed similar behaviour by triggering the specimen with a local external magnetic field. The velocities of the normal-resistive boundaries are very sensitive to the amount of heat generated behind the boundary in the normal region. At higher current amplitudes, or even at I_c for temperatures below the λ-point, Fig. 8(b), the time delay becomes very small.

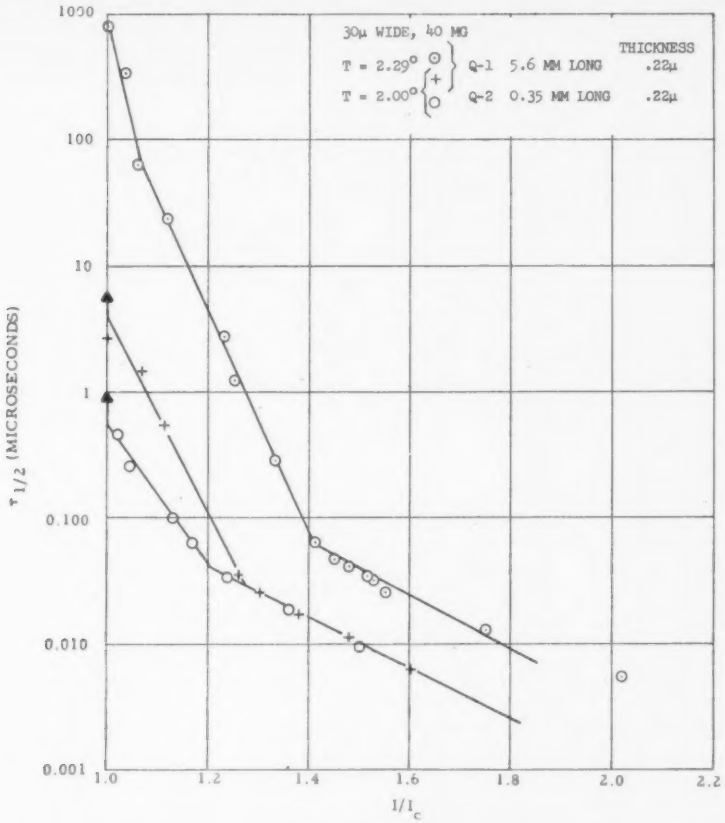


FIG. 9. Superconducting-to-normal transition times as a function of current-pulse amplitude, expressed as I/I_c.

This is evidently due to the greater power input. At still higher pulse amplitudes, the resistance is precipitated at more places along the length of the film, presumably at imperfections, until finally the linear segments can no longer be resolved and, for all practical purposes, the film switches simultaneously along its length. The pulse amplitude at which this occurs is sharply defined and is called I_s. Oscilloscope traces 8(c–e) show a typical behaviour as I_s is approached. For ease of comparison, the voltage sensitivity was adjusted so that the current and voltage amplitudes coincide when the film is fully switched. Both the voltage and current sensitivities are constant in all three traces so that the relative magnitudes of the current can be estimated from the scale.

For representing the dependence of transition time on drive current, bath temperature and film geometry, the transition time $\tau_{1/2}$, has been defined as the time delay between the current and voltage traces at half amplitude, corrected for the time required for the signal to make a return trip from the specimen. Transition time vs. I/I_c curves for specimen Q-1 at two different bath temperatures are shown in Fig. 9. Similar curves have been obtained for a variety of bath temperatures and appeared to establish no regular pattern.[4] However, it was discovered that if the normalizing current was chosen by extrapolating the I_c vs. T curve from below the λ-point to above the λ-point according to equation (5), the transition time vs. current curves then fell in an orderly pattern. Furthermore, they superposed in the region of fast switching. Thus fast transition times can be

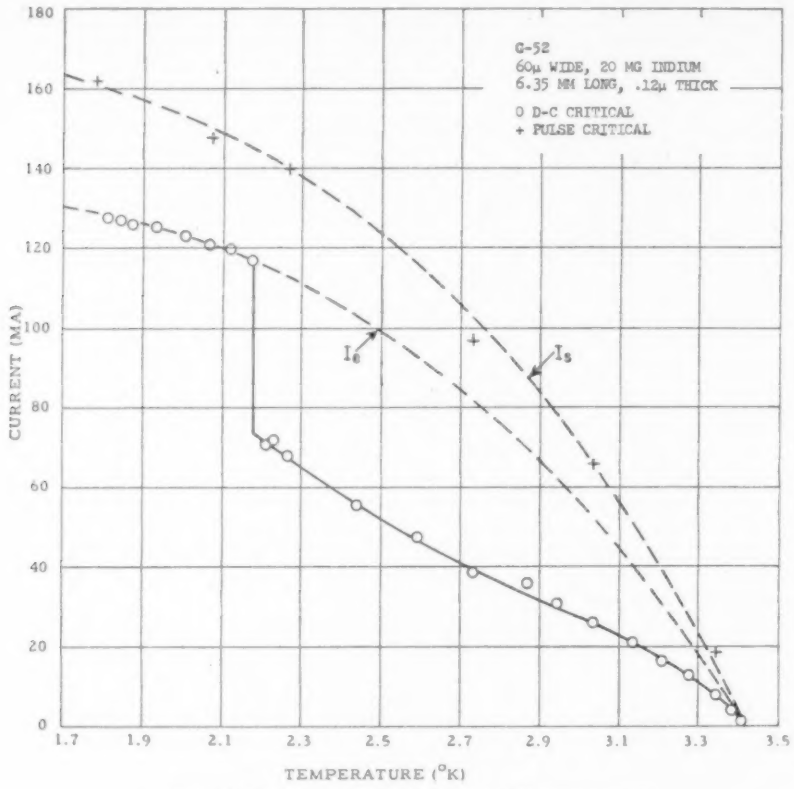


FIG. 10. Variation of critical current with temperature.

expressed as a function of I/I_e alone, where I_e is the extrapolated current. Fig. 10 shows a plot of I_e vs. T along with the fast switching current I_s vs. T. The ratio of I_s to I_e is constant and very nearly 1·25 for all specimens. In Fig. 9 transition-time curves for two specimens of different length are compared. In the fast switching region the two curves coincide; but, in the region where the normal phase propagates along the length of the film, the curves diverge as expected.

Transition times for specimens of various widths are compared in Fig. 11. In the length-dependent switching region no particular consistency can be expected since the transition might be initiated at several points. In the fast switching region, however, a regular pattern is observable. Apart from two anomalies which are explained later, the switching time is independent of width within experimental errors.

Transition times for specimens of various thickness are compared in Fig. 12. Fortunately, there are no anomalies and a regular dependence on thickness is clearly distinguishable. The dependence on thickness is shown in Fig. 13 for $I/I_c = 1{\cdot}25$.

Finally, it is of interest to compare transition times for geometrically identical specimens on different substrates. The measured results for a typical pair are shown in Fig. 14. At first sight the transition of a film on glass appears to take longer, but if the pulse currents for the two are normalized to the same I_c, namely that of quartz, the transition of the specimen on quartz takes longer. The effect, however, is not large and might easily be attributed to errors. To be certain, a number of pairs were compared and in nine cases out of ten the transitions on glass were faster when the normalizing current for a film on glass

was taken as the expected value for a geometrical identical specimen on quartz. Actually, in order to resolve the effect it was necessary in many cases to correct for width and thickness using equation (2).

As in the steady-state case, many of the above features can be understood by making an order-of-magnitude analysis of the heat flow. Of particular interest is the fast-switching region. Here only a very short time interval need be considered, which means that the problem can be treated as one-dimensional (normal to the surface of the film). Moreover, the heat flowing from the film directly into the bath in time intervals of interest has been estimated and shown to be negligible. Hence, for a substrate of thermal diffusivity D, the heat in time Δt will penetrate a depth $\Delta d = (D\Delta t)^{1/2}$. For Δt varying from 10^{-7} to 10^{-8} sec, Δd varies from approximately $1{\cdot}5\times10^{-4}$ cm to 5×10^{-4} cm for glass and from approximately 10^{-2} to 3×10^{-2} cm for quartz. This justifies the use of a one-dimensional analysis, especially for glass. Now if C_m and C_s are the heat capacities per unit volume of the film and substrate respectively, then the temperature rise of the film can be estimated from

$$(C_m\delta + C_s\Delta d)\Delta T = q\Delta t \qquad (6)$$

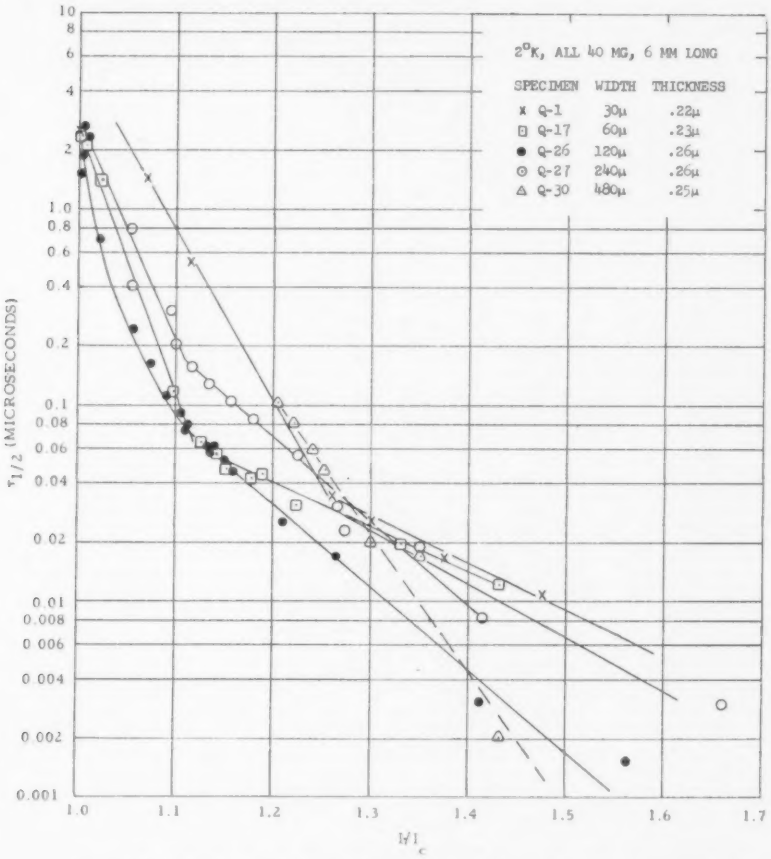


FIG. 11. Superconducting-to-normal transition times vs. current-pulse amplitude, expressed as I/I_c, for varying film widths.

For specimen G-106 (Fig. 3) at 3°K,

$$q = \frac{I^2 R_n r}{W \times \text{length}} \sim r\ (\text{W/cm}^2)$$

$$r = R/R_n, \quad R_n = 1{\cdot}5\ \Omega$$

$$C_m \delta = 3 \times 10^{-3} \times 2{\cdot}5 \times 10^{-5} \sim 7 \times 10^{-8}\ \text{J/cm}^2\text{-°K}$$

$$C_s \sim 4 \times 10^{-4}\ \text{J/cm}^3\text{-°K}$$

$$C_s \Delta d \sim 12 \times 10^{-8}\ \text{J/cm}^2\text{-°K, for } \Delta t = 5 \times 10^{-8}\text{sec}$$

Hence

$$\Delta T \sim 2{\cdot}5\, r \times 10^{-1}\ \text{°K}$$

From the r (I, T) curve for the continuous portion of the transition of the geometrically similar specimen on quartz (Q-17, Fig. 3), r changed by 0·01 while T changed by 0·1°K. The initial value of r was at least 0·01. This yields a change in ΔT the order of 2×10^{-3} °K, which is significant but not enough to change r appreciably. On the other hand, r is probably increasing at least an order of magnitude faster than the above in the fast-switching region, since the transition is initiated at more points. Now, if r starts at 0·1 and increases at a rate 10 times faster than observed below I_c, then $r = 1$ could easily be regeneratively attained in 5×10^{-8} sec. Actually, a more sophisticated treatment of this problem is required, but this crude analysis shows that the bulk of the transition could easily be due to heat, provided r starts

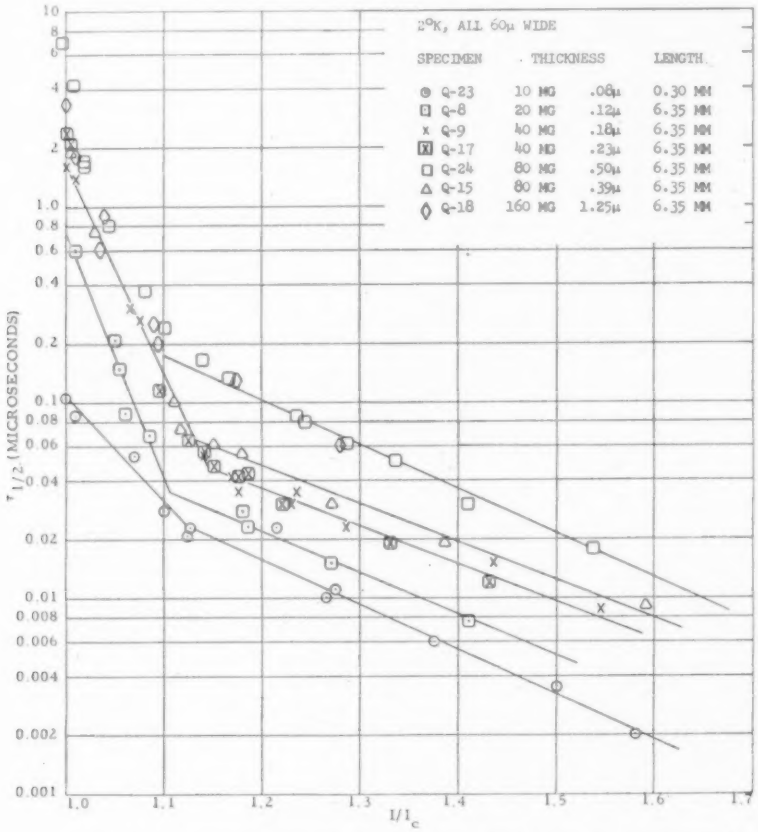

2°K, ALL 60μ WIDE
SPECIMEN THICKNESS LENGTH
Q-23 10 MΩ .08μ 0.30 MM
Q-8 20 MΩ .12μ 6.35 MM
Q-9 40 MΩ .18μ 6.35 MM
Q-17 40 MΩ .23μ 6.35 MM
Q-24 80 MΩ .50μ 6.35 MM
Q-15 80 MΩ .39μ 6.35 MM
Q-18 160 MΩ 1.25μ 6.35 MM
τ1/2 (MICROSECONDS)
I/Ic

FIG. 12. Superconducting-to-normal transition time vs. current-pulse amplitude, expressed as I/I_c, for varying film thickness.

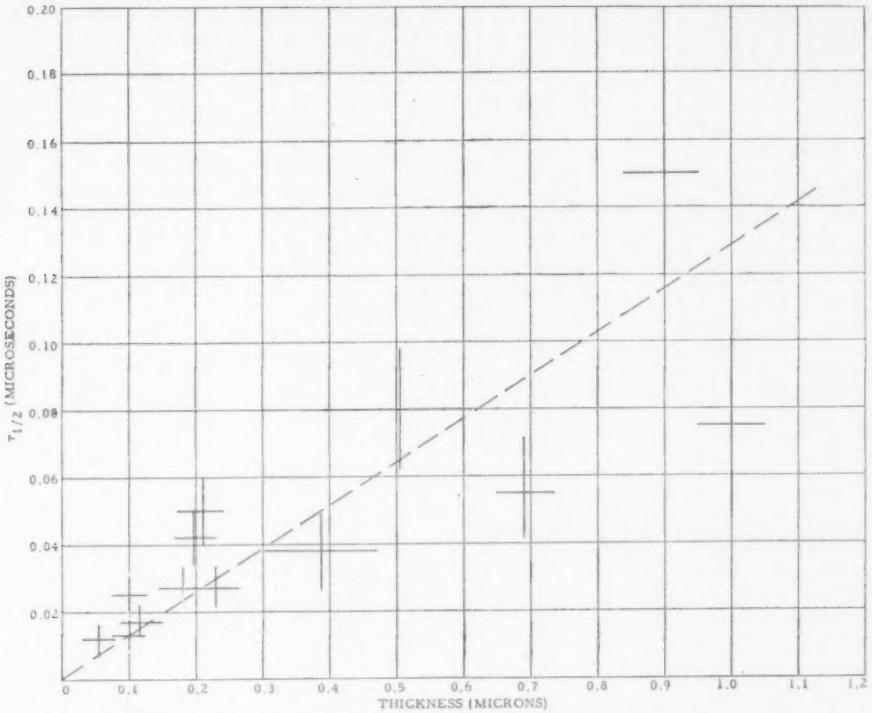


FIG. 13. Superconducting-to-normal transition time vs. film thickness for current pulses of amplitude 1·25 I_c. Errors indicated are only reasonable estimates.

increasing with I at a rate the order of 10 times that observed in the continuous portion of the transition.

It is now instructive to solve equation (6) for Δt in order to obtain the dependence of transition time on current and film thickness. Substitution of the expressions given earlier for q, etc., into equation (6), yields (after a little algebra)

$$\Delta t = \delta^2\left[\tfrac{1}{2}\alpha^2\left(\frac{I}{I_c}\right)^{-4}+\beta\left(\frac{I}{I_c}\right)^{-2}+ \tfrac{1}{2}\alpha\left(\frac{I}{I_c}\right)^{-2}\sqrt{\left[\alpha^2\left(\frac{I}{I_c}\right)^{-4}+4\beta\left(\frac{I}{I_c}\right)^{-2}\right]}\right] \tag{7}$$

where

$$\alpha = \frac{\sqrt{(k_s C_s)}\Delta T f(l'/\delta)}{l_0^2\rho_0 l' r}$$

$$\beta = \frac{C_m \Delta T f(l'/\delta)}{i_0^2\rho_0 l' r}$$

and k_s is the thermal conductivity of the substrate.

The origin of the terms in equation (7) can easily be traced by first setting $C_m = 0$ and then $C_s = 0$. In the first case (bulk of the heat absorbed in substrate),

$$\Delta t_1 = \alpha^2\left(\frac{I}{I_c}\right)^{-4}\delta^2, \text{ for } C_m = 0,$$

and in the second case (bulk of the heat absorbed in film),

$$\Delta t_2 = \beta\left(\frac{I}{I_c}\right)^{-2}\delta^2, \text{ for } C_s = 0.$$

Previous examination of the terms in equation (6) showed that the two could easily be contributing equally. Depending on the relative contributions of the C_m and C_s terms, log Δt vs. I/I_c, for $1<I/I_c<1{\cdot}5$, would have a slope ranging from -2 to -4. In either case Δt is proportional to δ^2. All the data presented above fall in this region. The data are certainly not good enough to resolve a δ from δ^2 dependence. The majority of the log $\tau_{1/2}$ vs. I/I_c curves shown in Figs. 11–14 have a

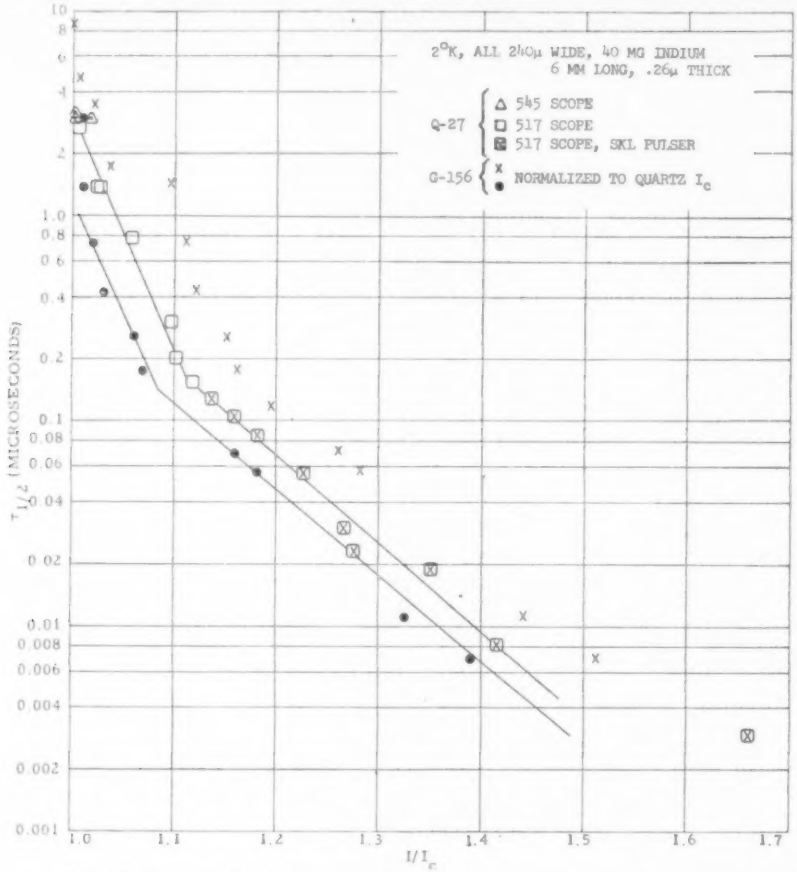


FIG. 14. Superconducting-to-normal transition times vs. current-pulse amplitude, expressed as I/I_c, for films differing only in substrate material.

slope of approximately $-2{\cdot}3$. A more detailed analysis could easily yield this value. Perhaps the most significant conclusion to be drawn from this analysis is that specimens on quartz should have characteristics profoundly different from those on glass. Their near identity indicates that the surface layer of quartz is probably glassy at least to a depth the order of 1μ. This could have easily resulted from optical polishing. The anomalous specimens Q-26 and Q-30 in Fig. 11 are now understandable in terms of heat "leaking through" the surface barrier.

Acknowledgements—The authors wish to thank Dr. LEE HEFLINGER of the Physical Research Laboratory of Space Technology Laboratories, Inc. for many stimulating and enlightening discussions. They also wish to thank Mr. CHARLES MARINOFF and Mr. SPENCER SPRIGGS for their assistance in preparing the test specimens.

REFERENCES

1. E. H. SONDHEIMER, *Advanc. Phys.* **1**, 1 (1952).
2. P. GRASSMANN and A. KARAGOUNIS, Heat transfer to boiling helium in the range of nuclear and film boiling, *Proceedings of the Fifth International Conference on Low Temperature Physics and Chemistry, Madison, Wisconsin, August* 26–30, (1957). University of Wisconsin Press (1958).
3. J. W. BREMMER and V. L. NEWHOUSE, *Phys. Rev. Letters* **1**, 15 (1958).
4. F. W. SCHMIDLIN and E. C. CRITTENDEN, Resistance changes in phase transition in superconducting thin film and fine wires, *Proceedings of the Fifth International Conference on Low Temperature Physics and Chemistry, Madison, Wisconsin, August* 26–30, 1957. University of Wisconsin Press 1958.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 335–338. Printed in Great Britain

HIGH-FREQUENCY BEHAVIOR OF VERY THIN SUPERCONDUCTING FILMS and CHANGES IN CRITICAL TEMPERATURE PRODUCED BY ELECTROSTATIC CHARGING*

R. E. GLOVER, III

University of North Carolina, Chapel Hill, North Carolina

A. HIGH-FREQUENCY BEHAVIOR

AT AUDIO frequencies a metal in the superconducting state behaves as if it were a perfect conductor. Energy can be given to the electrons by means of an electric field, but this energy is not passed on to the lattice. The loss mechanisms which result in joule heating for a metal in the normal state are inoperative. At the frequencies of visible light, however, the situation is different. Here the behavior of a metal in the superconducting state is indistinguishable from that in the normal state. The mechanisms by which energy given to the electrons results in heating of the lattice are operative in the superconducting as well as the normal material.

Measurements of the optical transmission of long-wave infrared radiation through very thin lead and tin films have shown that the transition between these two types of behavior takes place in the millimeter and submillimeter wave range.[1,2] Loss mechanisms first become effective in a superconductor at a critical frequency ω_g. They then increase rapidly as the frequency is further raised until the behavior of a normal metal is approached. This critical frequency is pictured as being due to a "gap" in the electron density of states of a superconductor.[3] At frequencies below ω_g the energy of the photons associated with the radiation are too small to excite electrons across the gap and the radiation cannot be absorbed. At frequencies above ω_g, however, the photon energies are sufficient to excite electrons across the gap and absorption becomes possible. For sufficiently high frequencies the width of the gap is insignificant compared with the photon energies, and in this case there is no difference in the behavior of a metal in the normal and superconducting states.

Measurements of the transmission of millimeter and submillimeter radiation through thin films have recently been greatly improved in accuracy and extended to include indium and mercury.[4] The wavelength range between 1·1 and 0·1 mm was covered. Aside from some small discrepancies the results are found to be in agreement with the predictions of the BARDEEN *et al.*[5] theory of superconductivity. The discrepancies between experiments and the theory may prove to be of considerable theoretical interest. They are, however, small enough to be neglected for most practical purposes.

Calculations based on the BARDEEN *et al.* theory lead to an expression for a complex conductivity $\sigma = \sigma_1 - i\sigma_2$ applicable in the case of thin films.[6] The expressions are derived for the 0°K limit but should be good approximations at temperatures as high as $1/2T_c$. The ratios of the real and imaginary parts to σ_N, the conductivity in the normal state, are plotted as a function of frequency in Fig. 1. The imaginary conductivity σ_2 describes the component of current that is 90° out of phase with the driving electric field while σ_1 is related to the in-phase current. Joule heating is associated only with σ_1. Given these functions it is immediately possible to calculate the reflection,

* Supported in part by the Office of Naval Research.

transmission and absorption coefficients of any superconducting film. This has been done in Fig. 2 for a film having a resistance per square of

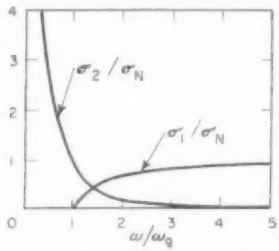


FIG. 1. Frequency dependence at 0°K of the effective complex conductivity for a thin film $\sigma = \sigma_1 - i\sigma_2$ as derived from the BARDEEN *et al.* theory. σ_N = conductivity of film in normal state; ω_g = energy gap frequency.

188 Ω (= $Z_0/2$). The frequency-independent coefficients for the film in the normal state are also given. A considerable degree of control can be exercised over the characteristics by properly choosing the normal-state resistance per square of the film. Nothing need be known about the thickness of the film except that it is thin compared to the wavelength of the radiation, the normal-state skin depth and the superconducting penetration depth. These conditions are usually fulfilled for films showing appreciable transmission in the normal state*. For the calculations it was assumed that the film was freely suspended. This is not necessary. The film could be deposited on any suitable transparent substrate. Reflection, transmission and absorption coefficients could again be calculated but would depend upon the index of refraction and thickness of the particular substrate chosen. A frequency scale reduced in terms of the gap frequency ω_g is given in the diagram. The corresponding wavelengths for a tin film are given at the top of the drawing.

It is seen in the figure that the optical properties of a film in the normal and superconducting states are almost the same for frequencies five times the gap frequency ω_g. At lower frequencies, however, the behavior of the film is quite different in the two states. For example, the transmission of a tin film to 4-mm radiation would be 2·5 per cent in the superconducting state and 25 per cent in the normal state, a change by a factor of 10. For 3-cm

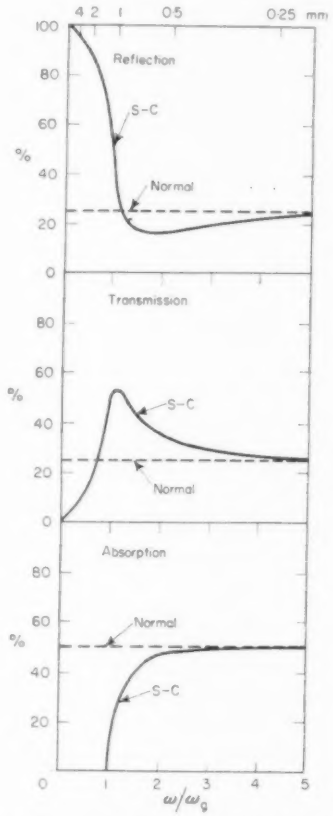


FIG. 2. Frequency dependence of the reflection, transmission and absorption coefficients of a film, as calculated from the functions in Fig. 1. The film is assumed to be freely suspended and to have a normal state resistance of 188 Ω/square. The wavelength scale shown at the upper edge is for a tin film having a transition temperature $T_c = 2{\cdot}73$°K and a gap energy of $3{\cdot}6\,kT_c$. S–C = superconductivity.

radiation the difference in transmission would be a factor of several hundred. This ratio could be further increased by using material having a higher transition temperature than tin.

A thin film acts as a microwave switch when

* The values of the resistivity obtained from measurements on bulk samples suggest that such a film would be less than an angstrom thick. However, because of surface scattering and the high degree of lattice disorder present in evaporated films, the film resistivity at 4°K is several orders of magnitude larger than the bulk value. This makes it possible to use films having thicknesses of the order of 30 Å.

it is changed from one state to the other. This change in state can be brought about by running a current through the film. Alternatively, a magnetic field could be applied. Relatively high fields would be needed, however, since the critical field increases with decreasing thickness. An intriguing possibility is that of switching the film by means of currents built up by incident microwave radiation. It might prove possible to switch the film rapidly enough so that it would be in the normal state during part of the cycle and in the superconducting state for the remainder. Because of the difference in transmission in the two states the film would operate as a microwave-frequency chopper. Such an arrangement should prove to be an effective generator of harmonics in the millimeter and submillimeter range.

We have seen, then, that the difference in transmission of a thin film in the normal and superconducting states could be used to modulate the flow of high-frequency radiation. This modulation should be very effective in the centimeter microwave region and moderately effective in the millimeter and submillimeter range. The change in reflection coefficient could, of course, also be used. In addition to modulation, the critical frequency dependence of reflection, transmission and absorption of thin superconducting films might make them useful for filtering and detecting radiation in the submillimeter range.

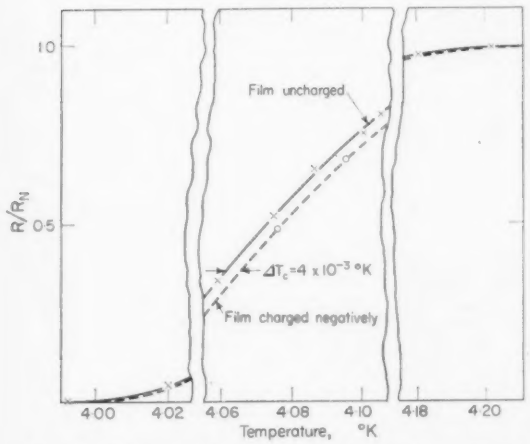


FIG. 3. Change in the transition of a tin film caused by negative charging. Approximately two extra electrons per hundred thousand atoms have been added. Film thickness = 90 Å; resistance/square at 4·2°K = 18 Ω $\equiv R_N$.

B. EFFECTS OF ELECTROSTATIC CHARGING

When a parallel-plate condenser is charged, electrons are added to one plate and removed from the other. For plates 0·01 mm thick the number of electrons that can be added before the best dielectrics break down is insignificant compared to the number of atoms in the plate. If, however, a film of the order of 50 Å thick is used as one plate of the condenser, it proves possible to add or subtract about one electron for each 10,000 atoms. It might be expected that this change in the carrier concentration would change the conductivity of the film. Such an effect has often been looked for over the last fifty years. Recent measurements in Germany[7] and Italy[8] have clearly shown that the normal state conductivity of metals can be altered by electrostatic charging. The relative change in conductivity is the same order of magnitude as the ratio of added electrons to the number of atoms present.

We have recently started studying this effect of electrostatic charging on the conductivity of thin metal films. In the case of gold films, adding electrons by negative charging increases the conductivity while positive charging reduces it. In the case of tin, we find that negative charging reduces the conductivity while positive charging increases it. For both gold and tin the magnitude of the relative conductivity change is the same for positive and negative charging.

Recently, measurements have been made to determine the effect of electrostatic charging on the superconducting transition temperature of tin films.[9] For all of the six films that have been tested so far, negative charging has increased the transition temperature while positive charging decreases it. Results for one film are shown in Fig. 3.

In the first part of the experiment the transition curve of resistance vs. temperature is measured with the film in the uncharged condition. Then at various points in the transition region the change of resistance upon charging is measured. These give the circles in Fig. 3 through which the displaced transition curve for the charged film is drawn.

Only the shift of the transition temperature to higher temperatures upon negative charging is

shown in Fig. 3. Based on a rough approximation of the film thickness, about two electrons have been added by charging for each hundred thousand atoms. This results in an increase in the transition temperature of 4 millidegrees. Positive charging reduced the transition temperature but only by about 1 millidegree. It is possible that this asymmetry is real. However, it can probably be explained on the basis of the non-uniform nature of the film. The transition curve was obtained by measuring the resistance of the film. Upon negative charging it is probable that the transition temperature is increased more in some regions than in others. The measuring current picks out the regions having the highest transition temperatures and the maximum effect is therefore observed. Upon positive charging the transition temperature is reduced. In this case the current flows through the regions whose transition temperatures are affected least, and the effect appears too small.

The fact that the measured transition temperature is appreciably lowered for positive charging suggests that the charged film cannot be thought of as a normal film with a surface charge concentrated within a very small distance of the surface. On such a picture the current would flow through the essentially unchanged region of the film and there would be no measured decrease in transition temperature.

REFERENCES

1. R. E. GLOVER, III and M. TINKHAM, *Phys. Rev.* **104,** 844 (1956).
2. R. E. GLOVER, III and M. TINKHAM, *Phys. Rev.* **108,** 243 (1957).
3. M. TINKHAM, *Phys. Rev.* **104,** 845 (1956).
4. D. M. GINSBERG and M. TINKHAM, *Phys. Rev.* **118,** 990 (1960).
5. J. BARDEEN, L. N. COOPER and J. R. SCHRIEFFER, *Phys. Rev.* **108,** 1175 (1957).
6. D. C. MATTIS and J. BARDEEN, *Phys. Rev.* **111,** 412 (1958).
7. A. DEUBNER and K. RAMBKE, *Ann. Phys. Lpz.*, **17,** 317 (1956).
8. G. BONFIGLIOLI, E. COEN and R. MALVANO, *Phys. Rev.* **101,** 1281 (1956).
9. R. E. GLOVER, III and M. SHERRILL, *Phys. Rev. Letters,* **5,** 248 (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, p. 339. Printed in Great Britain

RESEARCH ON SUPERCONDUCTIVE DEVICES IN SWEDEN

R. E. JACOBSSON

Chalmers University of Technology, Göteborg, Sweden

AT THE Royal Institute of Technology, Stockholm, Professor J. O. LINDE of the Department of Physics has a well-equipped low-temperature laboratory. It is built up around an ADL-Collins helium cryostat, which was installed in 1956 and which has recently been provided with arrangements for the production of liquid hydrogen.

At that laboratory, research is done on the electrical resistivity of noble-metal alloys and also on the magnetic properties of paramagnetic substances. In addition, this laboratory has hitherto been the only supplier of liquid helium for other low-temperature research in this country.

Among the users of liquid helium is the Research Laboratory of Electronics at the Chalmers University of Technology in Göteborg, where a maser programme is being carried out. Two solid-state masers have been put in operation which are intended for research in radio astronomy.

Work on superconductive devices has been carried out in Sweden for about a year. The objective is to investigate the possibilities of using superconductive circuitry in computing systems. Two groups of people with background in circuit techniques have been formed, one at the Royal Institute of Technology, the other at the Chalmers University of Technology.

The work is still in a preparatory phase. At Chalmers we have built up our low-temperature apparatus, including a cryostat for measurements in the helium range. We are engaged in the study of the technological problems of obtaining thin-film structures and we intend to install a unit for vacuum evaporation at 10^{-9} mm Hg.

P. BENGTSSON and H. LINDE of the Royal Institute of Technology have built a vacuum-evaporation unit for the production on a laboratory scale of thin-film structures.

The vacuum chamber, which is made of aluminum, is pumped down with an oil-diffusion pump (Edwards F203). The pressure obtained after degassing (ionic bombardment) and with a liquid-air trap is 5×10^{-7} mm Hg measured with an ionization vacuum-meter. During evaporation the vacuum is better than 5×10^{-6} mm Hg. The work chamber is equipped with a revolving device for changing masks and evaporation sources under vacuum.

The masks are prepared in the conventional manner by etching a copper-foil coated with Kodak Photo Resist in a pattern determined by a photographic negative exposed on the resist. An edge sharpness of 5–10 μ is attainable in this way.

The equipment described has been used to evaporate silver, lead, tin and silicon monoxide to form cryotron matrices, with an evaporation speed of approximately 100 Å/sec.

An automatic tracer to provide plots of resistance versus current in thin films has been built from a potentiometer recorder.

To date, preliminary measurements of the static properties of some films have been carried out. Systematic studies of the static and dynamic behaviour of superconducting thin films are planned.

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 340–342. Printed in Great Britain

SUPERCONDUCTING FILMS LESS THAN 100 Å THICK*

DAVID ABRAHAM

Sperry Gyroscope Company, Great Neck, New York

THE research described herein had been initially motivated by a desire to determine if superconducting elements might provide the basis for ultra-high-speed computers. One of the principal speed limitations of cryotrons being produced in 1957 was their very low resistance when in the normal state. It was hoped that, by depositing extremely thin superconducting films of high resistance, the ratio of the inductance to the normal-state resistance might be lowered appreciably, making possible film cryotrons with very fast switching speeds. In the course of depositing fiims it was found that films of tin having a thickness of less than 20 Å and a normal resistance of more than about 200Ω/square could not be deposited reproducibly or preserved for any considerable length of time. However, just within this reliability threshold, films have a resistivity which suggested that they would naturally lend themselves for use as non-reflective elements in high-frequency transmission lines when in the normal state. When in the superconducting state the films should reflect energy as would good bulk conductors.

Two classes of experiments seemed to be called for: (A)—A study of the critical-current dependence on film configuration of very thin films. (B)—The behavior of films of the above type in response to electrical signals containing frequency components well into the microwave region.

A. CRITICAL CURRENT DEPENDENCE ON FILM CONFIGURATION

The basic mechanism associated with switching induced by critical currents has never been clearly resolved. For bulk superconductors, Silsbee hypothesized that a critical current was that current whose self-magnetic field was the critical magnetic field for the specimen under test. This hypothesis has been adequately verified for bulk superconductors. However, experiments with films of thicknesses of the same order of magnitude as the London penetration depth, or thinner, have shown anomalous results. H. London suggested in 1935 that the kinetic energy associated with the electron drift of the critical current might be the basic mechanism causing switching in very thin films.[1] Several experiments have been designed in an attempt to resolve the question of how films 20–100 Å thick can switch under the influence of a current of critical value. These experiments are: (1) a determination of the critical-current dependence on film width, (2) critical-current dependence on film thickness and (3) an attempt at a direct verification of Silsbee's hypothesis. In the event that the third experiment should yield negative results it would then become necessary to perform additional experiments.

All d.c. experiments were performed using thin films of tin since the transition between states for tin is quite sharp. An additional reason for this choice was that detailed studies of the superconducting behavior of tin films could be made at and near the critical temperature without the difficulties one might encounter by using such materials as lead. An even more impelling motivation was the desire for the continued use of a material with which great familiarity had been developed in earlier experiments. In every instance films were deposited, in a vacuum ranging from 5×10^{-5} to 5×10^{-7} mm of Hg, on crystalline-quartz substrates. These films were coated with SiO to protect them from the test environment because equipment was not yet available to test their superconducting properties in the vacuum in which they were deposited. The low-temperature tests were performed with the film and substrate completely immersed in a liquid-helium bath, the

* This research was supported primarily by the U.S. Navy Bureau of Ships, under contracts NObsr 72728 and NObsr 77521.

bath temperature controllable to within 10^{-3}°K over a temperature ranging from 2°K to 4·5°K.

Sets of five films of graduated widths were deposited simultaneously on a single substrate in an attempt to insure that the films be identical in all other respects. In prior experiments, depositions of five films of identical width were performed as a control. These experiments showed that in most cases the five films had variations in critical currents of less than 5 per cent, variations in normal-state resistance of less than 15 per cent, and critical temperatures ranging within a few hundredths of a degree Kelvin. The experiments with films of graduated widths resulted in an approximately linear relationship between critical current and film width for the films in each set tested. The deviations from a linear relationship were not systematic so that there was no reason to assume that any particular correction factor should be applied to the results. The general relationship cannot be given explicitly because the relative thickness of each set of films is not known with any degree of precision. The thickness determination was based on the measurement of the temperature dependence of film conductivity. For films as thin as those deposited, the applicability of this method of determination of thickness is open to serious doubt.

The result of the foregoing experiment is similar to that reported by CRITTENDEN at the Conference on Thin Films, at Bolton Landing, New York, for films of about 500 Å thickness. It renders very plausible the notion that the critical-current distribution is uniform and that the current density has a critical value. It is much more difficult to conceive of a current distribution giving rise to a critical self-magnetic field requiring that the total current be directly proportional to film width.

If the critical-current dependence on film thickness should prove to be linear for films which are thin as compared with the London penetration depth, it would strengthen arguments favoring a critical current-density interpretation, although it would in no way weaken any arguments favoring Silsbee's hypothesis. The performance of such an experiment awaits the completion of a constant-evaporation-rate metallic-vapor source and its calibration. Upon the completion of this source, five films, identical except for thickness, will be deposited by moving a shutter in such a way as to expose the five film areas to the source for graduated time periods.

An experiment which may directly validate or invalidate the applicability of Silsbee's hypothesis for very thin films will be performed. The apparatus has been constructed which will make possible the deposition of two superposed films separated by a thin (1000 Å) dielectric layer of silicon monoxide. For a critical current which is distributed over a portion of film width large as compared with the separation between films, the magnetic field associated with this distribution will be almost the same in each film. By increasing the current in one film slightly above the critical value, the other film should be subjected to a critical field if Silsbee's hypothesis holds. If switching is not observed, however, it will be necessary to determine if the current distribution meets the above criterion. If it does, it will be safe to state that Silsbee's hypothesis does not hold for films thin as compared with the penetration depth. If the current distribution in the first film is exceedingly irregular, and the second film does not switch, no conclusion can be safely drawn.

B. THE RESPONSE OF SUPERCONDUCTING THIN FILMS TO ELECTRICAL SIGNALS HAVING FREQUENCY COMPONENTS WELL INTO THE MICROWAVE REGION

It would be interesting to determine if any inherent limitations in the switching time of thin superconducting films exist. A superficial inquiry into the nature of superconductivity leads one to believe that no limitation longer than the order of 10^{-13} sec should exist. The basis for this belief is the quantum-mechanical nature of the transitions that electrons undergo during the phase change. The availability of films having a resistivity of about 200 Ω/square makes it possible to determine if films can undergo transitions at frequencies up to approximately 50 kMc/s. The basis for this claim lies in the fact that the ratio of the superconducting to the normal microwave transmission (and inferentially the resistivity) is significantly less than one for frequencies as high as 50 kMc/s, depending on the material of which the film is composed.[2] If the film will undergo transitions under the influence of either the electric or magnetic fields associated with a microwave signal, the microwave transmission and reflection properties of the

film would change significantly. This change would give rise to harmonic generation which can be detected.

The speed of transition from the superconducting to the normal state was measured as being less than $1{\cdot}5 \times 10^{-10}$ sec by employing a scheme dependent on the change in resistance of a film going through a transition. The experiment consisted of a fast-rising square-wave d.c. pulse propagated down a coaxial transmission line, with a coaxial stub terminated by a film of the line's characteristic impedance when in the normally conducting state inserted in the line, to an Edgerton, Germeshausen and Grier travelling-wave oscilloscope with a rise time of $1{\cdot}5 \times 10^{-10}$ sec. When the pulse travelling down the stub was of less than the critical value, the superconducting film reflected the pulse with opposite polarity cutting short the pulse viewed on the oscilloscope screen. Increasing the pulse amplitude above the critical value caused the film to switch to the normal state with a resultant loss of reflection, the viewed pulse rising back to its original value as soon as the transition was completed. As viewed with the oscilloscope, this rise was limited by the rise time of the scope indicating that very fast transitions were being realized in the laboratory. The cycle rate in this experiment was limited to 120 c/s by a mercury wetted switch which formed the basis of the pulser. The experiment was not performed in such a way as to make a measurement of maximum transition time from the normal to the superconducting state possible.

An experiment was performed to determine if a superconducting film could be used as a microwave gate. A film was placed entirely across an *X*-band waveguide and a 17-dB attenuation in power transmission was observed as the film underwent a change from the normal to the superconducting state.

With the above background, an experiment was designed to determine if a superconducting film could switch repetitively at 9 kMc/s under the influence of a microwave field. A test fixture was built, in which the film would be utilized as a probe in a waveguide to coaxial transition section. The film would be oriented in such a way as to be exposed entirely to the peak values of the microwave field in the guide and to have the field across its width of almost uniform value. If switching were to take place, the coupling of energy from the waveguide into the coaxial line would change radically four times each cycle, giving rise to substantial second-harmonic generation which would be detected at the output of the coaxial line. This experiment was performed only once as of April 1960. The microwave power available was limited to 1 W and no switching was observed. Arrangements are currently being made to provide much greater peak power levels using bursts of microwave energy available from a pulsed magnetron source.

If the foregoing experiment should yield positive results, it would then be repeated at higher frequencies. The possibility of producing relatively inexpensive, broadband, cryogenic, microwave devices operating in the *K*-band frequency range hinges on the success of such experiments. If this possibility materializes, exceedingly fast microwave computers of reasonable bulk may become economically feasible.

REFERENCES

1. D. Shoenberg, *Superconductivity* pp. 173–178. Cambridge University Press (1952).
2. R. E. Glover, III and M. Tinkham, *Phys. Rev.* **108**, 248 (1957).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 343–350. Printed in Great Britain

CONTINUOUS-SHEET SUPERCONDUCTIVE MEMORY

L. L. BURNS, G. W. LECK, G. A. ALPHONSE and R. W. KATZ

RCA Laboratories, Princeton, New Jersey

INTRODUCTION

PERSISTENT-CURRENT superconducting memory cells for computer applications are well known and have been reported in the literature.[1–3] A type of persistent-current memory cell worked on at the RCA Laboratories is illustrated in Fig. 1. A cross-sectional view of this memory cell is

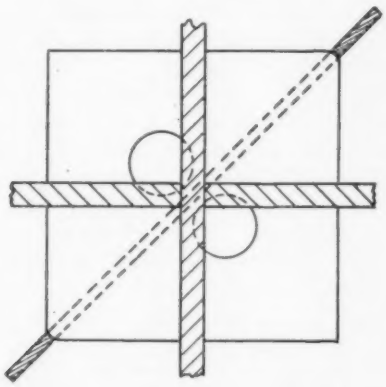


FIG. 1. Persistent-current memory cell.

shown in Fig. 2, and Fig. 3 illustrates the operation: as the current in the drive pulse increases a field is reached that causes the center portion of the cell to become normal and the flux lines then link the sense line causing a sense output signal as

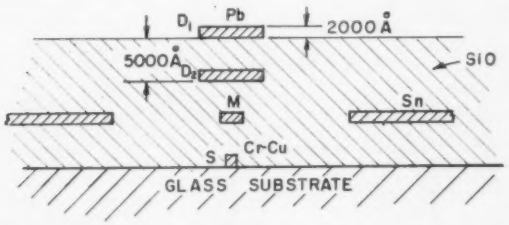


FIG. 2. Cross-section of persistent-current memory cell.

shown. As the drive current decays towards zero a persistent current is set-up circulating the two holes and a trapped flux exists as illustrated. The material used for the superconducting film in this type of memory is tin, and very little joule heating is obtained during the transition. Essentially ideal superconducting memory-cell operation is therefore obtained. The advantages of this type of cell are the same as those of other persistent-current cells of this type: x, y selection can be used with its advantages for computer memory systems, fabrication is simple, and very small sizes in large arrays are possible.

CYCLE AND STORAGE TIME IN TWO-HOLE MEMORY

In order to operate the cell as a reliable computer element certain conditions must be imposed on the magnitude, duration and rise-time of the selecting and interrogating pulses. The characteristics of these pulses together with the resistivity of the memory material in its normal state and its physical configuration determine such features as the level of the persistent current and the storage and cycle time of the memory. A typical cycle of operation is shown in Fig. 4. This shows that initially a setting pulse is used to store a "0" into the cell. Then a drive current of magnitude between 0·5 I_c and I_c (0·6 was chosen in this work) is sufficient to operate the cell if the pulse duration is longer than the thermal recovery time of the memory. This operation is possible because the drive current adds up to the persistent current to make the "bridge" go to the normal state, and the magnetic field of the drive can link the holes. The drive current is kept until the bridge has returned to the superconducting state, and the existing field remains trapped. This process stores a persistent current of magnitude equal to that of the drive current. A reversal of the polarity of the

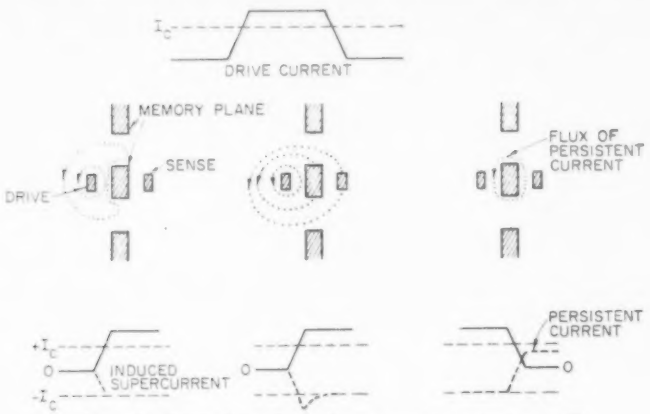


FIG. 3. Operation of persistent-current memory cell.

drive current will reverse the direction of the persistent current. In Fig. 4 the values of 0·3 I_c show that the half-select currents do not disturb the stored information. A photograph of one of the early single cells is shown in Fig. 5, and an array of 6×1 cells designed for the study of the reproducibility problem is pictured in Fig. 6.

It was just pointed out that in order to store a persistent current in the two-hole memory it is necessary to bring the bridge of the cell into the normal state and let it return to the superconducting state with an applied external field. The

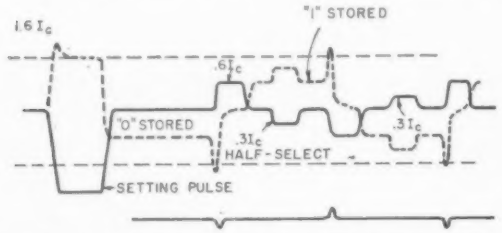


FIG. 4. Read–write cycle for a cryoelectric memory cell.

time for this process to take place determines the storage time of the memory. This time consists of two parts: the fundamental switching time for the phase transition of the material, and the time (thermal-recovery time) taken by the material to heat up then cool down to a temperature below the critical value T_c. The fundamental switching time is of the order of 10^{-9} sec for a lead film 1000 Å thick. It is rather small compared to the thermal recovery time which is the main factor governing the storage time and the cycle time of the memory. It is difficult to get an exact expression for this thermal recovery time (also called thermal relaxation time). An approximate evaluation can be obtained from the following considerations. Let T be the temperature of the material whose thermal recovery time is desired, T_c its critical temperature and T_0 the temperature of the surrounding helium bath. Here it is assumed that the material is in direct contact with the helium bath. The normal current $i_n(t)$ raises the temperature of the specimen from T_0 according to the relation

$$\int R i_n^2(t)\, dt = MC(T - T_0) \tag{1}$$

where M is the mass of the portion of the material which has become normal, C is its specific heat and R is the resistance of the material in the normal state. The time derivative of equation (1) gives the rate of heat generation

$$\text{rate}_1 = \frac{R}{MC} i_n^2(t) \tag{2}$$

The material cools down to T_0 at the rate

$$\text{rate}_2 = k(T - T_0) \tag{3}$$

where k is a constant in sec^{-1}. Then the heat equation of the material is

$$\text{rate}_1 - \text{rate}_2 = \frac{\partial T}{\partial t} = \frac{R}{MC} i_n^2(t) - k(T - T_0) \tag{4}$$

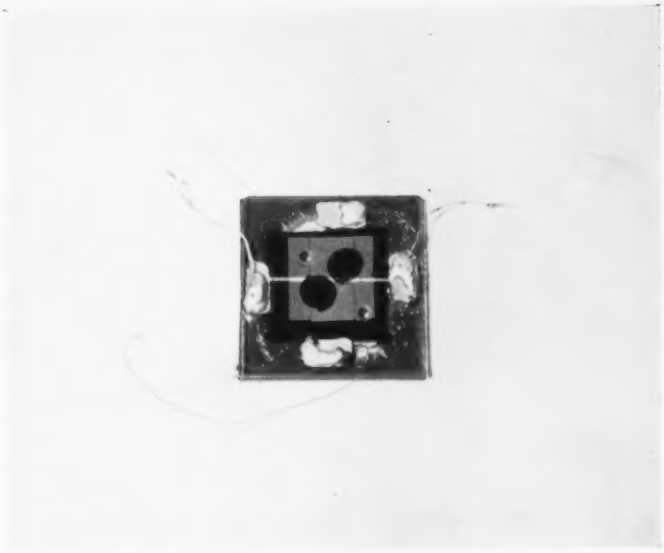


FIG. 5. Single experimental memory cell.

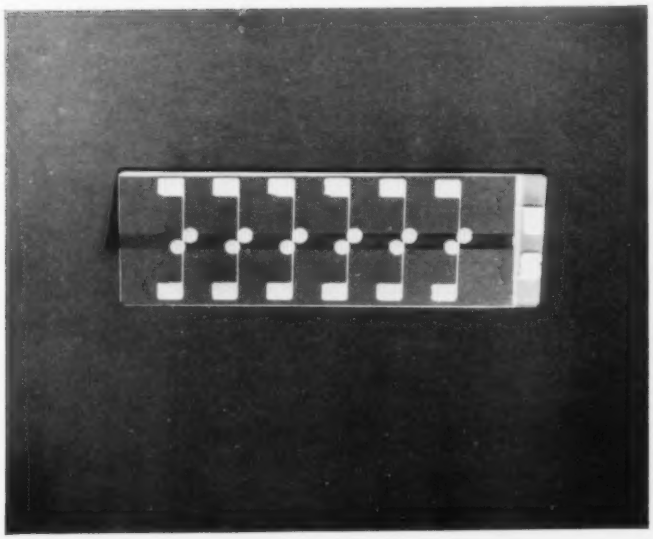


FIG. 6. Six-cell array.

facing p. 344

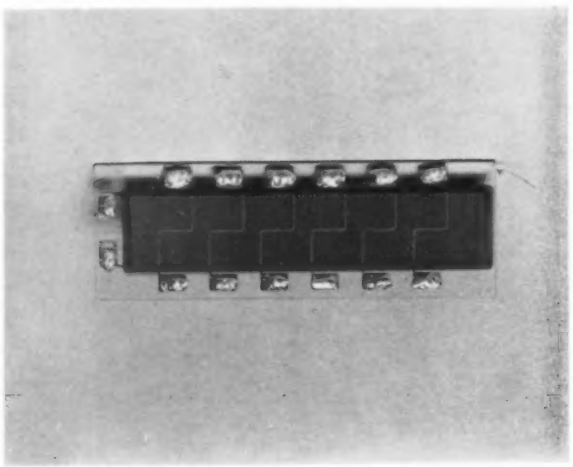


FIG. 8. Six-cell continuous-sheet array.

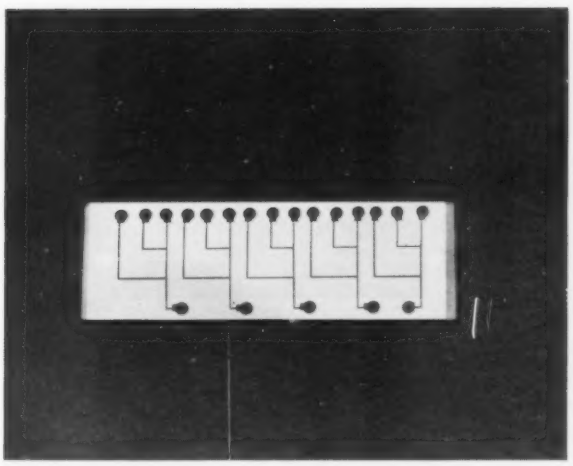


FIG. 15. Five-cryotron test arrangement.

After time t, one gets by integration

$$T-T_0 = \frac{R}{MC}\int_0^t i_n^2(t)\,dt - k\int_0^t (T-T_0)\,dt \qquad (5)$$

The material reaches its maximum temperature at a time t_1 such that $(\partial T/\partial t)_{t=t_1} = 0$. Thus, from

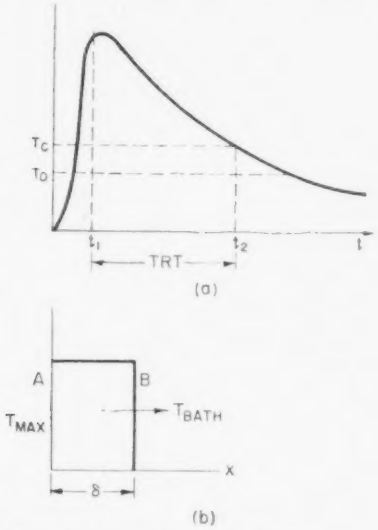


FIG. 7. Thermal recovery time.

equation (4) the maximum temperature is reached when

$$T_{max}-T_0 = \frac{R}{kMC} I_1^2 \qquad (6)$$

where $I_1 = i_n(t_1)$. The time t_1 is of the order of 10^{-8} sec for a tin cell. The curve of equation (5) is sketched in Fig. 7(a). In order to find the thermal recovery time (TRT) one finds the time t_2 needed for the material to return to the temperature T_c. Then

$$\text{TRT} = t_2 - t_1 \qquad (7)$$

This is done by considering the cooling rate [equation (3)] of the specimen (the bridge in the memory cell)

$$\frac{\partial T}{\partial t} = k(T-T_0)$$

which yields

$$T-T_0 = A\,e^{-kt} \qquad (8)$$

The constant A is obtained from the condition

$$T = T_{max} \text{ at } t = t_1$$

or

$$T_{max}-T_0 = A\,e^{-kt_1} \simeq A$$

since $t_1 \simeq 0$

Thus, from equation (6),

$$A \simeq \frac{R}{kMC} I_1^2 \qquad (9)$$

At time t_2 the temperature reaches T_c; thus

$$T_c-T_0 = \frac{R}{kMC} I_1^2\, e^{-kt_2}$$

and the thermal recovery time simplifies to

$$t_2 = \ln\left[\frac{kMC(T_c-T_0)}{RI_1^2}\right]^{-1/k} \text{sec} \qquad (10)$$

after neglecting t_1. The constant k in the expression for the thermal recovery time is to be found by measuring the temperature at any time along the curve of Fig. 7(a).

If there is a silicon monoxide insulation layer between the element and the helium bath one can get an approximate solution for the heat relaxation time by solving the one-dimensional flow equation

$$\frac{\partial^2 T}{\partial X^2} = \frac{1}{\alpha 2}\frac{\partial T}{\partial t} \qquad (11)$$

for an SiO slab of thickness X, and finding the time required for A to reach the temperature T_0, Fig. 7(b). In this equation, α is the thermal diffusivity of SiO.

Drive pulses for persistent-current memory cells of the type described depend on the film thicknesses used, but 300-mA pulses are typical. Sensing technique to avoid stray pick-up from the drive line must be carefully worked out. The most satisfactory arrangement is to use a completely balanced sense line and individual amplifiers in each line followed by differential arrangement; with this technique stray pick-up from the drive lines can be reduced to less than 0·1 mV. Sense outputs

from cells of this type are about 10 mV with insulation and superconducting films of 2000Å thickness. A cell switching time of less than 80 mμ sec was calculated. Measured switching time was 60 mμ sec.

The one limitation on actual application of cells of the type described is reproducibility. Typically, rather wide variations are obtained from cells of this type and it has been determined experimentally that this is mostly due to edge effects along the bridge that actually switch from normal to superconducting during operation. To eliminate the undesirable effects of edges on the reproducibility problem, a construction involving no holes or other discontinuities in the memory film has been used. Fig. 8 illustrates a six-cell arrangement used for studying the reproducibility problem on cells of this type. A maximum variation of 10 per cent is obtained on cells of this type and by careful processing much lower values of maximum variation can be obtained. The place of the physical holes that were present in the previously described cells has now been replaced by normal regions punched in the superconducting film by the drive currents. Thus, actual holes are not necessary and fabrication is greatly simplified.

INFORMATION STORAGE IN CONTINUOUS PLANES

A continuous superconducting memory plane stores information as a persistent current at the boundary between the normal and superconducting phases, the normal phase being enclosed by the superconducting phase. In the equilibrium

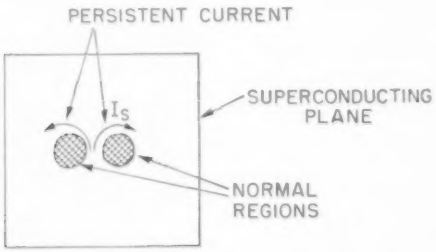


FIG. 9. Normal regions in a continuous-sheet cell.

position the "trapped flux" is sufficient to keep the enclosed portion in the normal state. The principle of storage of persistent currents in such a memory plane is the same as for the multiply connected memory but the holes are replaced by normally conducting regions. A typical cell with a persistent current in it is shown in Fig. 9.

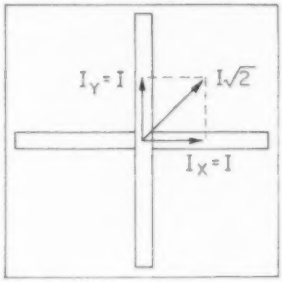


FIG. 10. Drive current relationship.

Fig. 10 shows how a point in the plane can be selected by X and Y drives for memory operation. The drive currents I_x and I_y are of the same magnitude I. Along the surface of the memory plane the magnetic field of the drive has a value

$$H_0 = \frac{I}{2\pi\delta} \tag{12}$$

where δ is the average distance between the drive line and the plane (it is approximately the same for both drive lines). The fields of the two drives add up vectorially to give a tangential component of $H = H_0\sqrt{2}$ and a normal component equal to $2H_0$ as shown in Fig. 11. The result is that the points P and P' are the first to become normal. The rest of the operation is illustrated in Fig. 12 and is very similar to the process described in Fig. 3. The only difference is that the normal regions start at P and P', then spread to a radius $r = PQ$ according to the peak value of the applied field, and the area underneath the drive wires becomes normal. When the drive current drops to a value such that $r < w/2$, the portion of the plane near point O' returns to the superconducting state and persistent current remains after the removal of the external field.

From geometric considerations it is possible to determine the optimum number of cells per memory plane of a certain dimension. (Fig. 13 shows a cell with the drive wire shrunk to a point to facilitate the analysis.) Referring to Fig. 14,

one sees that the maximum diameter of each cell must be less than the distance between two consecutive cells if destructive interaction is to be avoided. Let the distance between any two consecutive cells be a and let there be $N = n \times n$ cells per

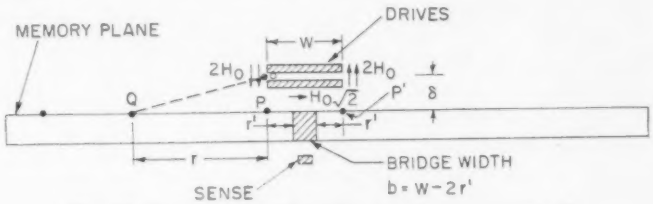


FIG. 11. Cross-section of continuous-sheet memory cell.

square plane of side A. From the previous argument, the condition for no interference between cells is such that

$$r \leqslant a/2 \tag{13}$$

FIG. 12. Action of flux pattern during write operation.

Since $A = na$ or $A^2 = n^2a^2 = Na^2$, it follows that the area of the plane is governed by the inequality

$$A^2 \geqslant N(r^2) \tag{14}$$

This means that the total number of cells N per memory plane is such that

$$N \leqslant \frac{A^2}{4r^2} \tag{15}$$

The radius r is related to the field at point Q in Fig. 13 by the fact that the applied field must

FIG. 13. Geometry of continuous cell.

have the critical value at point Q if the normal region is to extend to that point. Thus, at point Q one has

$$H_c = \frac{I}{2\pi\rho} \text{ or } \rho = \frac{I}{2\pi H_c} \tag{16}$$

and

$$r^2 = \rho^2 - \delta^2 \simeq \rho^2 \tag{17}$$

Thus substituting (16) into (17) and (17) into (15) one gets

$$N \leqslant \frac{A^2}{I^2/\pi^2 H_c^2} \tag{18}$$

Equation (18) gives the optimum number of cells for a memory plane of side A. It also determines the theoretical upper limit on the drive current if a certain number N of cells per memory plane is specified.

It was stated that the actual radius r' of the cell is less than the value r which occurred during the transition. The value of r' is determined by the equilibrium condition between the normal and the superconducting phases. We also know that it

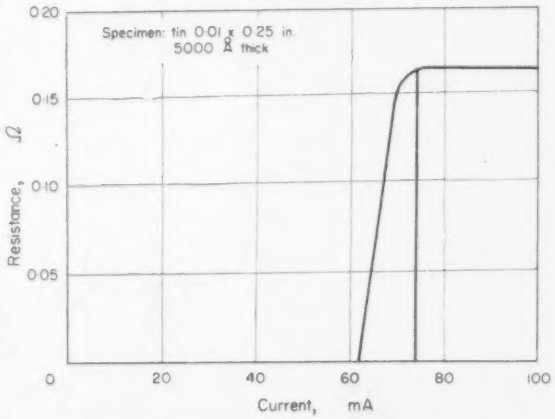


FIG. 14. Cryotron gate self-switching curve.

is less than half the width of the drive line, i.e. $r'_{max} = w/2$. The energy associated with a persistent current of density J_s at the boundary of the normal-to-superconduction region is related to the difference between the internal free energies per cubic meter of the two phases by the equilibrium condition (in e.s.u.)

$$f_n - f_s = \tfrac{1}{2}\lambda J_s^2 \tag{19}$$

where λ is a constant depending only on temperature and on the material. This energy is stored in the field H_s of the super-current by the relation

$$f_n - f_s = \tfrac{1}{2}\mu H_s^2 \tag{20}$$

It follows that

$$\mu H_s^2 = \lambda J_s^2 \tag{21}$$

or

$$\mu \frac{I_s^2}{(2\pi r')^2} = \lambda J_s^2 \tag{22}$$

and

$$r' = \frac{1}{2\pi}\sqrt{\left(\frac{\mu}{\lambda}\right)}\frac{I_s}{J_s} \tag{23}$$

Since the width of the bridge is

$$b = w - 2r' \tag{24}$$

one can thus determine the conditions on the magnitude of the persistent current for a desired bridge dimension. The current density can be obtained from the relationship

$$J_s = \frac{1}{\sqrt{\lambda}} H_s e^{-\beta z} \tag{25}$$

where

$$\beta^2 = \frac{1}{c^2\lambda^2} \tag{26}$$

The values of z are a measure of the penetration inside the superconducting region.

It should be noted that the transition occurs with the minimum generation of joule heat because the screening current exists only in the superconducting phase where there is no electric field in the direction of the current. This is especially true in a purely crystalline substance. For a polycrystalline substance (actual practice) there may be a component of electric field in the direction of the current, thus accounting for a small joule effect. At any rate the power loss is less than in the two-hole memory cell. Although the losses are at a minimum, there is a certain amount of energy required from the driving source during the transition. It manifests itself as a change in free energy and serves to maintain the persistent current at the boundary. That a certain reversible (reversible means not lost in the form of heat) amount of energy is necessary in the information storage process can be verified from the following considerations on the equilibrium of energy at the boundary between a superconducting and a normal region. For two vectors A and B, the following relation holds:

$$\boldsymbol{B}\cdot(\nabla\times\boldsymbol{A}) - \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B}) = \nabla\cdot(\boldsymbol{A}\times\boldsymbol{B}) \tag{27}$$

Applying this relation to the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ and recalling that

$$\nabla\times\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \qquad \nabla\times\boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$$
$$\boldsymbol{B} = \mu\boldsymbol{H}, \qquad \boldsymbol{D} = \epsilon\boldsymbol{E} \tag{28}$$

and that for the superconductor (the subscripts o and s stand for the normal and superconducting electrons)

$$\boldsymbol{J} = \boldsymbol{J}_o + \boldsymbol{J}_s, \qquad \boldsymbol{J}_o = \sigma\boldsymbol{E}$$
$$\boldsymbol{G} = \lambda\boldsymbol{J}_s, \qquad \frac{\partial \boldsymbol{G}}{\partial t} = \boldsymbol{E} \tag{29}$$

one gets for the Poynting vector:

$$\boldsymbol{H}\cdot(\nabla\times\boldsymbol{E}) - \boldsymbol{E}\cdot(\nabla\times\boldsymbol{H}) = \nabla\cdot(\boldsymbol{E}\times\boldsymbol{H}) \tag{30}$$

or

$$\boldsymbol{H}\cdot\left(-\frac{\partial}{\partial t}\mu\boldsymbol{H}\right) - \boldsymbol{E}\cdot\left(\boldsymbol{J}_o + \boldsymbol{J}_s + \frac{\partial}{\partial t}\epsilon\boldsymbol{E}\right)$$
$$= \nabla\cdot(\boldsymbol{E}\times\boldsymbol{H})$$

or

$$\boldsymbol{H}\cdot\left(-\frac{\partial}{\partial t}\mu\boldsymbol{H}\right) - \boldsymbol{E}\cdot\left(\sigma\boldsymbol{E} + \frac{\partial}{\partial t}\epsilon\boldsymbol{E}\right) -$$
$$\boldsymbol{J}_s\cdot\frac{\partial}{\partial_t}\lambda\boldsymbol{J}_s = \nabla\cdot(\boldsymbol{E}\times\boldsymbol{H})$$

This simplifies to

$$\nabla\cdot(\boldsymbol{E}\times\boldsymbol{H}) + \frac{\partial}{\partial t}(\epsilon E^2 + \mu H^2 + \lambda J_s^2) = -\sigma E^2 \tag{31}$$

It was pointed out that the right-hand side of the equation (representing the joule heating) is small

or negligible. Thus the continuous-plane memory is basically a faster device than the two-hole memory. This is true in the sense that no normal current needs to flow in the path of the persistent current. The factor λJ_s^2 represents the reversible amount of energy required for the transition, or, more specifically, for the storage of information. It is positive or negative, depending on whether the transition is into the superconducting state or into the normal state. This kinetic-energy density of the supercurrent can be entirely recovered. This energy, as previously stated, is generated at the boundary between the two phases and is thus conserved as a free energy in connection with the phase transformation.

STORAGE TIME OF CONTINUOUS-SHEET MEMORY

The significance of the expression in equation (31) is that the divergence of the energy density vector $\boldsymbol{E}\times\boldsymbol{H}$ equals the decrease of the local energy density W per second (power density). Since $\sigma E^2 = 0$ in the superconductor, equation (31) reduces to

$$\nabla\cdot(\boldsymbol{E}\times\boldsymbol{H}) = -\frac{\partial W}{\partial t} \tag{32}$$

where $W = E^2 = \mu H^2 + \mu J_s^2$.
The application of Gauss' theorem yields

$$\int_V \frac{\partial W}{\partial t}\,dV = -\int_S (\boldsymbol{E}\times\boldsymbol{H})\cdot D\boldsymbol{S} \tag{33}$$

where V and S stand for volume and surface respectively. Because W has the discontinuity $\lambda/2\ J_s^2$ at the boundary, the velocity v of the boundary must be accounted for by writing according to LONDON[4]

$$\frac{d}{dt}\int_V W\,dV = \int_V \frac{\partial W}{\partial t}\,dV + \int_S \frac{\lambda}{2} J_s^2 \boldsymbol{v}\cdot d\boldsymbol{S} \tag{34}$$

The application of (34) to equation (33) yields

$$\frac{d}{dt}\int_V W\,dV = -\int_S (\boldsymbol{E}\times\boldsymbol{H})\cdot d\boldsymbol{S} + \int_S \frac{\lambda}{2} J_s^2 \boldsymbol{v}\cdot d\boldsymbol{S} \tag{35}$$

If the thickness of the material is a constant C, one can write $dV = C\,dS$, and equation (35) becomes

$$\frac{d}{dt}\int_V CW\,dS = -\int_S (\boldsymbol{E}\times\boldsymbol{H})\cdot d\boldsymbol{S} + \int_S \frac{\lambda}{2} J_s^2 \boldsymbol{v}\cdot d\boldsymbol{S} \tag{36}$$

Differentiating this expression with respect to the area S, one gets

$$\frac{d}{dt}(CW) = -|\boldsymbol{E}\times\boldsymbol{H}| + \frac{\lambda}{2} J_s^2\, v \tag{37}$$

and the velocity of the boundary is (in e.s.u.)

$$v = \frac{d/dt\,(CW) + |\boldsymbol{E}\times\boldsymbol{H}|}{\lambda/2\, J_s^2} \tag{38}$$

It was previously mentioned that during the storage process in the continuous-memory plane the radius of the normal region grows to a value r then reduces to a value r' less than $w/2$. The velocity of this boundary motion is now given by equation (38). The storage time thus consists of the time required by the boundary to move from 0 to r, then from r to r'. Since

$$dr = v\,dt \text{ or } dt = \frac{dr}{v} \tag{39}$$

it follows that the storage time τ is

$$\tau = \int_0^r \frac{dr}{v} + \int_r^{r'} \frac{dr}{v} \tag{40}$$

where the values of r and r' are given by equations (17) and (23). This time is obviously much shorter than the storage time in the two-hole memory because there is less heat dissipation.

CRYOTRONS

Thin-film evaporated cryotrons are ideal for the selection switches for cryoelectric memories. The obvious advantages include their complete compatibility, very low cost, and satisfactory speeds. Reproducibility is a major problem with cryotrons. A typical self-switching curve of the gate of a cryotron is shown in Fig. 14. Proper attention to the details of fabrication have resulted in this very sharp transition. One difficulty or limitation

ordinarily present in cryotrons is that of their slow speed, but this is not present in a selection tree for use in driving a cryoelectric memory; speed is important but gain is not necessary and cryotron switching in better than 100 mμsec with no gain can be readily attained. The non-stringent requirements for the cryotron in this application greatly reduces the restrictions normally placed on their limits of operation. A photograph of a set of five cryotrons used in reproducibility studies is shown in Fig. 15. The potential leads used for checking their characteristics are clearly visible.

FABRICATION

The fabrication of cryoelectric devices for actual applications assumes equal importance with theoretical considerations when all the advantages of cryoelectric devices are to be exploited. Various techniques can be used for fabricating cryoelectric devices, including: etched foil, plated films, and evaporated-film construction. Of these three techniques the evaporated-thin-film type of construction appears to be the most versatile and practical. The other techniques, such as plating, for example, offer certain advantages but one of the main disadvantages is the excessive handling required in the production of a single operational plane. This handling leads to many opportunities for errors and accidents, any one of which could completely destroy the plane. Contrasted with this, thin-film evaporation techniques, wherein the entire fabrication takes place within the vacuum chamber controlled by semi-automatic machinery, reduce the chances for human error to a minimum and handling to practically zero, resulting in a usable device practically every time.

Experimental memory cells of the type described have been made with a maximum variation in characteristics of less than 1 per cent between cells. To obtain this amount of reproducibility requires close attention to detail during the fabrication process.

CONCLUSIONS

A comparison has been made between persistent-current cryoelectric memory cells employing a multiply connected structure and those employing a simply connected structure. The comparison has shown that the latter structure is superior in speed and much simpler to make. Reproducibility between cells has been measured as better than $\frac{1}{4}$ per cent. This type of cell therefore offers considerable promise for computer applications.

REFERENCES

1. J. W. Crowe, *IBM J. Res. Dev.* **1,** 295 (1957).
2. R. L. Garwin, *IBM J. Res. Dev.* **1,** 304 (1957).
3. E. H. Rhoderick, *Brit. J. Appl. Phys.* **10,** 193 (1959).
4. Fritz London, *Superfluids* Vol. 1, pp. 137–141. Wiley, New York (1950).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 351–356. Printed in Great Britain

CHARACTERISTICS OF FILM CRYOTRONS

M. L. COHEN and J. L. MILES
Arthur D. Little, Inc., Cambridge, Mass.

A CRYOTRON is a switching device, much like a relay, in which one current controls another. The characteristics of importance therefore are what current is necessary to switch the cryotron, how much current can be switched, and how fast the device is or how much resistance is available for switching. One would also like to know how these parameters vary, what determines their values and how they are interrelated.

CRYOTRON MEASUREMENTS

In order to obtain the characteristics of film cryotrons and design parameters for cryotron circuits, single cryotrons are made and subjected to certain simple tests. For example, for each cryotron the normal gate resistance R_g and the critical gate current I_{gc}, as a function of temperature T, are measured. The critical controlling current I_{cc} is measured by the following technique. With an X–Y recorder, a graph of gate voltage vs. control current I_c is obtained for various values of gate current I_g. Fig. 1 shows a typical set of these curves, which we call the "quenching characteristics". The break in these curves, where the gate resistance increases rapidly, is considered to be the switching point, and an I_g vs. I_c plot of the switching points yields the "control characteristic" of the cryotron. This test is repeated at various temperatures, and the control characteristics are extrapolated to zero gate current in order to obtain the critical controlling currents at these temperatures, as shown in Fig. 2.

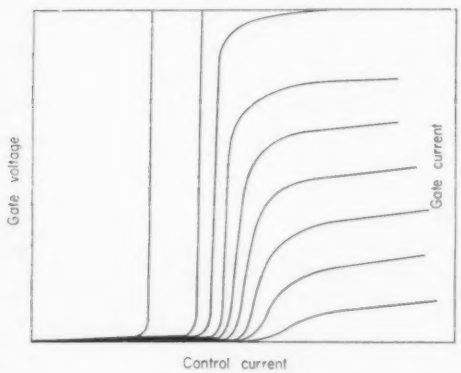


FIG. 1. Quenching characteristics.

FIG. 2. Control characteristics.

All these measurements are made on cryotrons of different gate thicknesses t, so that as the direct result of measurements on film cryotrons we have $I_{gc}(T,t)$, $I_{cc}(T,t)$ and $R_g(t)$.

COMPUTED CHARACTERISTICS

The measured characteristics apply to cryotrons of specific dimensions. For cryotron-circuit design more general parameters are desirable. Since, to a good first approximation, I_{gc} and I_{cc} are proportional to the gate and control widths, w_g and w_c, respectively[1]

$$J'_{gc} = \frac{I_{gc}}{w_g} \text{ and } J'_{cc} = \frac{I_{cc}}{w_c} \qquad (1)$$

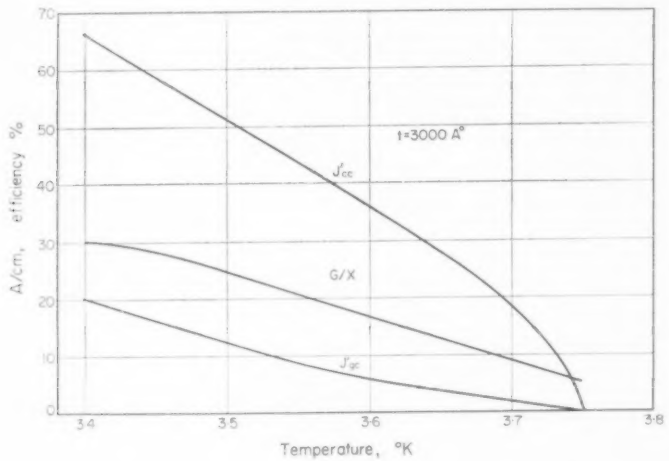


FIG. 3. Basic parameters vs. temperature.

are the desired parameters. The area resistivity ρ/t may also be computed:

$$\rho/t = XR_g \qquad (2)$$

where $X = w_g/w_c$ is the crossing ratio of the cryotron.

Finally, just as the current gain G of a cryotron is computed by dividing I_{gc} by I_{cc}, a general current gain, G/X may be computed by dividing J'_{gc} by J_{cc}. This last parameter may also be called "Silsbee efficiency"* or just "efficiency" because it also represents the ratio between internal critical field (that due to conducted current) and external critical field.

Thus, the basic cryotron parameters are: $J'_{gc}(T,t)$, $J'_{cc}(T,t)$, $G/X(T,t)$ and $\rho/t(t)$. The behavior of these functions is shown in Figs. 3 and 4.†

THE FIGURES OF MERIT, M_0, M_1, M_2

Now consider the problem of designing a cryotron circuit. The above parameters are functions of temperature and gate thickness. The first questions to be answered are naturally: What temperature and what gate thickness should be used? The efficiency G/X increases as the temperature is reduced; therefore, it is desirable to operate at as low a temperature as possible. The choice of thickness presents more of a problem. The area resistivity increases rapidly with decreasing thickness, but the Silsbee efficiency drops. The relative weight of these two opposing factors depends on the circuit in question.

Consider first a circuit in which a cryotron is driving only an identical cryotron. If all the interconnecting-lead inductance can be neglected, the time constant of this circuit is L_c/R_g. The inductance of the control is given by[1]

$$L_c = 4\pi \times 10^{-9}\, d\frac{l_c}{w_c} = k_1 X \text{ (henrys)} \qquad (3)$$

* From Silsbee's hypothesis which attributes quenching due to conducted current to the magnetic field associated with the current.

† These curves represent measured values and are typical of our "best" cryotrons.

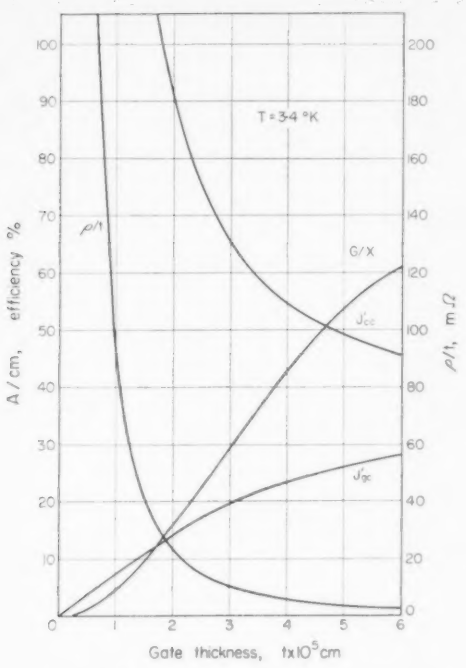


FIG. 4. Basic parameters vs. thickness.

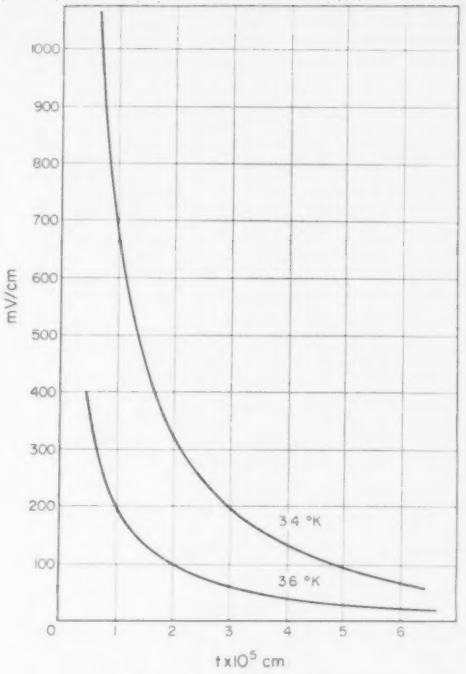


FIG. 5. M_0 vs. thickness.

and the gate resistance is

$$R_g = \frac{\rho/t}{X}$$

so that

$$L_c/R_g = \frac{k_1 X^2}{\rho/t} \tag{4}$$

For a given required current gain the crossing ratio is inversely proportional to the Silsbee efficiency; therefore

$$\frac{L_c}{R_g} \propto \frac{1}{\rho/t\left(\frac{G}{X}\right)^2} \tag{5}$$

The figure of merit M_2 is therefore defined by the equation

$$M_2 = \rho/t\left(\frac{G}{X}\right)^2 \quad \text{(ohms)} \tag{6}$$

High values of M_2 mean small L_c/R_g time constants and high speed in those circuits where the cryotron control inductance predominates.

Now consider a circuit in which the inductance of the cryotron controls is negligible compared to the inductance of the interconnecting leads; i.e. a circuit in which inductance is not proportional to the gate width. For a given current gain

$$\frac{L}{R} \propto \frac{1}{R_g} \propto \frac{1}{\rho/t\frac{G}{X}} \tag{7}$$

For this circuit the figure of merit

$$M_1 = \rho/t\frac{G}{X} \quad \text{(ohms)} \tag{8}$$

determines the optimum thickness.

Lastly, there are those circuits in which the output voltage available is the primary concern and there is no current-gain requirement. For these circuits we define a third figure of merit

$$M_0 = \rho/tJ'_{gc} \quad \text{(volts/cm gate length)} \tag{9}$$

These parameters, M_0, M_1 and M_2, are also

functions of temperature and gate thickness. Typical values are shown in Figs. 5, 6 and 7. Suitable gate thickness can be determined by

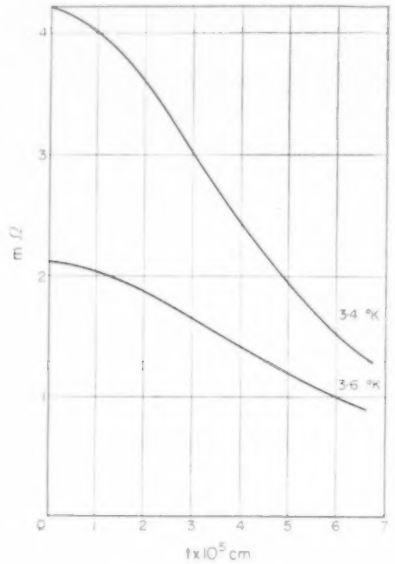


FIG. 6. M_1 vs. thickness.

noting the behaviour of the M-curve applicable to the particular circuit and temperature.

THE LOW-TEMPERATURE LIMIT FOR CRYOTRON CIRCUITS

As mentioned above it is usually desirable to operate cryotron circuits at as low a temperature as possible because of the increase in efficiency at low temperatures. However, at some temperature the external critical field for tin becomes greater than the internal critical field for lead. Cryotrons may not be operated below this temperature without the lead controls self-quenching before the tin gates are quenched by the controls. Furthermore, if the cryotron current gain is more than unity and the controls are required to carry as much current as the gates without quenching, the low-temperature limit will be higher, increasing with current gain, as shown in Fig. 8. At present our measurements of this limit are not exact, but it is clear that the fundamental limit is well below 3·0°K and that, for moderate values of current

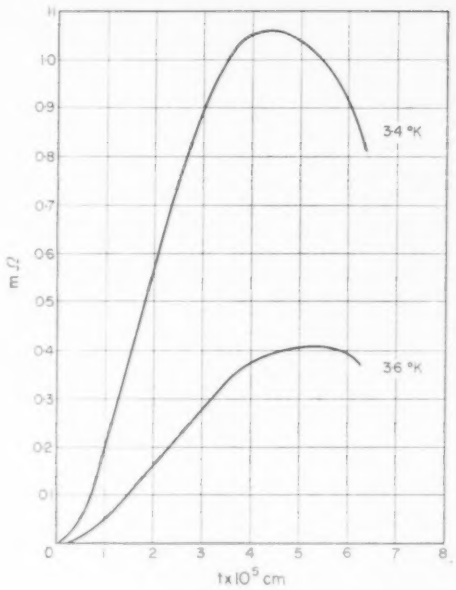


FIG. 7. M_2 vs. thickness.

gain (three to five), lead–tin cryotrons may be safely operated down to 3·4°K.

COMPARISON OF GATE MATERIALS

So far only the characteristics of lead–tin cryotrons have been discussed, and the various parameters have been considered to be functions of temperature and gate thickness alone. These parameters, of course, are also functions of gate material. One should therefore consider what might be gained through the use of gate materials other than tin. For example, gate resistivity can be increased by alloying. Will an alloy cryotron be better than a pure-tin cryotron? In general, those factors which increase resistivity, decrease efficiency. Gate thickness is a typical factor of this kind. Obviously, the M-parameters provide a basis for comparison of different materials just as they provide a basis for the comparison of gate thicknesses.

Some preliminary experiments have been performed with indium–tin alloys. The results are shown in Fig. 9. There is a notable increase in resistivity, but, as expected, the efficiency falls so that the net improvement shown by M_1 and M_2

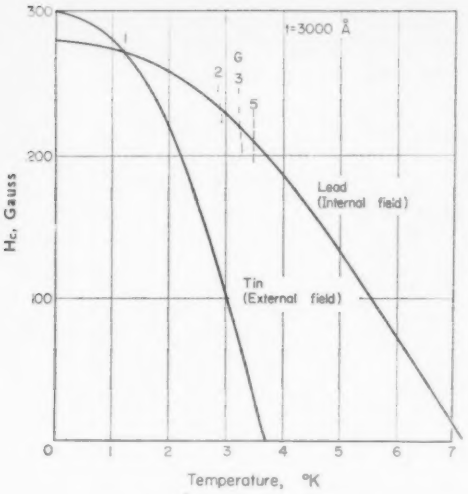


FIG. 8. Low-temperature limit.

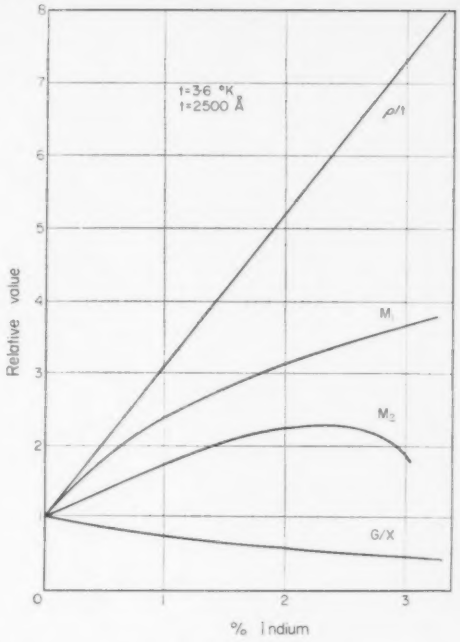


FIG. 9. Effect of indium.

are factors of 4 and 2, respectively. These results are for a 2500 Å gate thickness, which is not the optimum thickness for either parameter, and therefore they do not represent the maximum improvement obtainable with these alloys.

DESIGN OF CRYOTRON CIRCUITS

To illustrate another use of these parameters, consider the design of a three-stage cryotron ring oscillator. Since the minimum gain required is 2,[2] let us specify a current gain of 3 and an operating temperature at 3·4°K. An oscillator is nearly an M_2 type of circuit; therefore, a gate thickness of 3000 Å is close to optimum.* At this thickness and temperature the efficiency from Fig. 4 is 30 per cent, and the crossing ratio, therefore, must be 10.

The time constant of each stage may now be computed from the value of M_2 (Fig. 7, 0·9 mΩ) using the formula

$$L/R = \frac{k_1 G^2}{M_2} = \frac{10^{-12} \times 3^2}{9 \times 10^{-4}} = 10^{-8}\,(\text{sec}) \quad (10)$$

Actually, the time constant will be longer because of circuit inductance other than the cryotron control. If the time constant is assumed to be 2×10^{-8} sec, the maximum frequency of oscillation will be about 10–12 Mc/s.[2]

Finally, the control and gate widths may be determined by dividing the desired critical current by the appropriate critical-current density. For example, let I_{cc} be 1 A, J'_{cc} (from Fig. 4) is 65 A/cm, and w_c is therefore 0·015 cm or 0·006 in. And, since the crossing ratio is 10:1, the gate width is 0·06 in.

The oscillator described above has been built, and is shown in Fig. 10. When operated at temperatures of 3·4–3·45°K and supply currents of 1·75–2·0 A, it oscillated at frequencies of from 1 to 6 Mc/s. Table 1 compares the design and measured values for this oscillator.

CONCLUSION

From certain measurements on cryotrons, basic general parameters may be computed. Three figures of merit, the M parameters, are defined as certain ratios and combinations of the basic parameters. Although the M parameters, being derived, contain no new information, they are useful in cir-

* The M_2 peak lies between 4000 and 5000 Å. The selection of a 3000 Å gate thickness is conservative in that it provides a margin, 4, of safety against poor cryotrons at only a small cost in speed.

Table 1

	R_g (MΩ)	I_{cc} (A)	L/R (μsec)	f (Mc/s)
Design	1·0	1·0	20	10–12
Measured	0·85	~1·0	40–45*	6

* Computed from frequency.

cuit design and provide a basis for comparison and selection of cryotron gate materials and thicknesses to suit particular circuit applications.

REFERENCES

1. A. E. Slade, *Proc. I.R.E.* **48,** 1569 (1960).
2. M. L. Cohen, *Proc. I.R.E.* **48,** 1576 (1960).

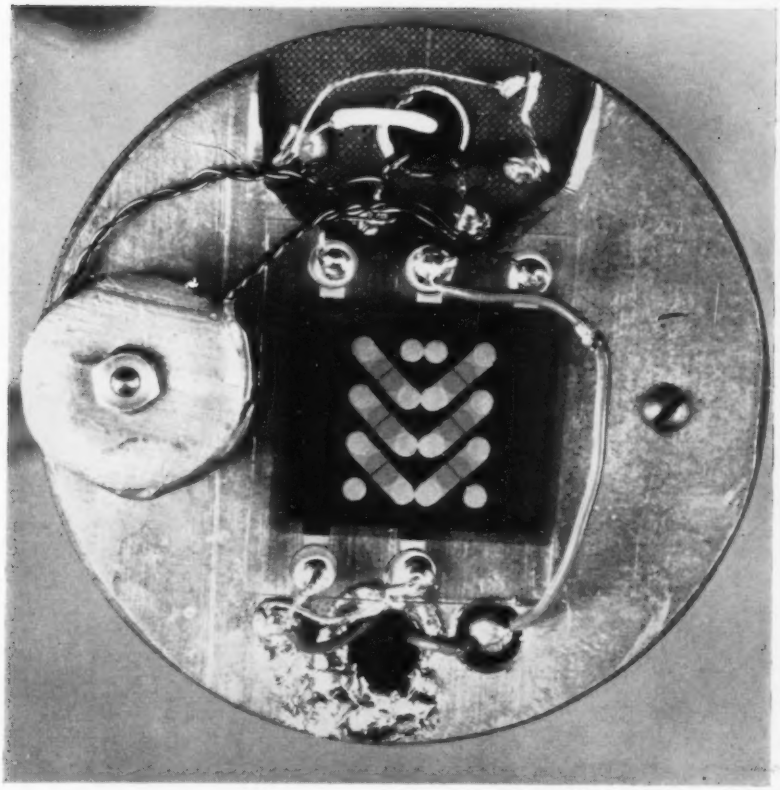


FIG. 10. Cryotron ring oscillator without transformer.

[*facing p.* 356

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 357–362. Printed in Great Britain

THIN-FILM CRYOTRON CATALOG MEMORY

A. E. SLADE and C. R. SMALLMAN

Arthur D. Little, Inc., Cambridge, Massachusetts

THE use of thin-film cryotrons in logical circuitry, such as flip-flops and shift registers, has already been discussed in previous literature.[1–3] The cryotrons used in this type of circuitry must be active devices with a greater-than-unity current gain. There are other applications, however, where low-gain cryotrons may be used. In switching networks and memories, the cryotrons can be driven from external transistor circuits and thus eliminate the necessity for the cryotrons to have gain greater than unity. As a result, considerably shorter switching times may be achieved.

Translators, persistent-current memories,[4–7] catalog memories[8] and storage and retrieval systems are excellent examples of this type of circuit. This paper will discuss the use of the low-gain cryotron in a catalog memory.

COMPARISON OF CRYOTRON TIME CONSTANTS

Let us first consider the time constant of high-gain cryotrons. Cryotron circuits in which a cryotron gate drives one or more cryotron controls require high-gain cryotrons. The time constant of a thin-film cryotron when used in this type of circuit is determined by the inductance of a control and the resistance of a gate:[9]

$$\frac{L_c}{R_g} = \frac{4\pi d l_c / w_c}{\rho l_g / t w_g} \times 10^{-9} \text{ sec} \qquad (1)$$

where l_c and l_g are the lengths of the control and gate, w_c and w_g are the widths of the control and gate, d is the distance from the control to the ground plane, t is the thickness of the gate and ρ is the low-temperature resistivity of the gate. Since $l_c = w_g$ and $l_g = w_c$

$$\frac{L_c}{R_g} = 4\pi \frac{dt}{\rho} \left(\frac{w_g}{w_c}\right)^2 \times 10^{-9} \text{ sec} \qquad (2)$$

Switching networks, wherein the cryotron controls are externally driven by high-impedance pulse generators and the cryotron gates do not drive other cryotron controls, do not require high-gain cryotrons. In this type of circuit, the time constant of a cryotron is determined by the inductance of a *gate* and the resistance of a gate rather than by the inductance of a *control* and the resistance of a gate. The time constant of a cryotron for this application is given by

$$\frac{L_g}{R_g} = 4\pi \frac{d l_g / w_g}{\rho l_g / t w_g} \times 10^{-9} \text{ sec} = 4\pi \frac{dt}{\rho} \times 10^{-9} \text{ sec} \qquad (3)$$

It can be seen that the time constant of the low-gain cryotron is independent of its length or width and is smaller than that given in equation (2) by a factor equal to the crossing ratio, w_g/w_c, squared.

The time constant for a high-gain cryotron with a crossing ratio of 10:1 is approximately 10^{-8} sec whereas the time constant of a low-gain cryotron in a switching network may be of the order of $2{\cdot}5 \times 10^{-11}$ sec. The factor d in equation (3) refers to the distance between the gate and the ground plane which may be a factor of 2 to 5 less than d in equation (2), depending upon the insulation thickness.

Device time constants such as these are useful in comparing the merits of different cryotron designs and for calculating circuit time constants. The time constant of a circuit will be greater than the device time constant depending upon the number of cryotrons involved and the compactness of the layout. A typical flip-flop, for example, has a circuit time constant of about $4\,L_c/R_g$. The time constant of a switching network will be longer than L_g/R_g by a factor determined by the number

of gates in series, and the layout. The time constant of a switching network will be discussed in greater detail in a later section of this paper.

THE CATALOG MEMORY

The catalog memory is an example of a circuit containing low-gain cryotrons. It is essentially a comparator where all of the words in a catalog can be compared simultaneously with any other word. The output of the memory is a yes or no answer as to the comparison.

In 1957 a small engineering model of this circuit was constructed in this laboratory with wire-wound cryotrons. It contained 25 eight-bit words and used 1802 cryotrons. It was found that not only was the memory difficult to construct with wire-wound devices but it was also very slow.

Development of thin-film cryotrons has prompted reconsideration of the catalog memory. A thin-film memory should be considerably easier to construct and in addition have a very short interrogate time.

DESCRIPTION OF CIRCUIT

A simplified circuit diagram of a thin-film catalog memory is shown in Fig. 1. It consists of

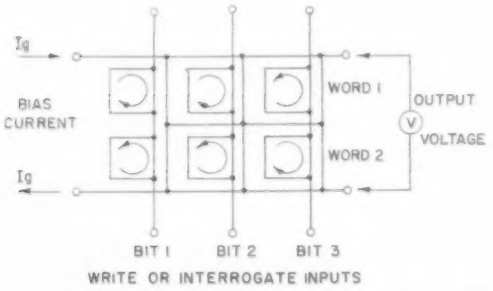


FIG. 1. Simplified circuit diagram of the catalog memory.

an output-gate network and persistent-current memory cells.

All bits of the memory are interrogated in parallel with either a positive or negative current and these interrogate currents either add to or subtract from the persistent current in the memory cells. When the interrogate current and the persistent current are additive, their combined magnetic field is sufficient to cause the corresponding gate in the gate network to become resistive.

If the interrogate currents match all of the persistent currents for one word of the memory, then all of the gates in that word become resistive and a voltage will appear at the memory output terminals. Unless there is an exact comparison there will always be at least one superconducting path through the memory and there will be no voltage output.

THE PERSISTENT-CURRENT MEMORY CELL

A number of persistent-current memory cells have been proposed[4–6] and many other configurations are possible. Basically they are all quite similar and it is a matter of choosing a cell configuration that fits the design of the memory in which it is to be used.

The memory cell used in the catalog memory is shown in Fig. 2. It consists of an enable cryotron

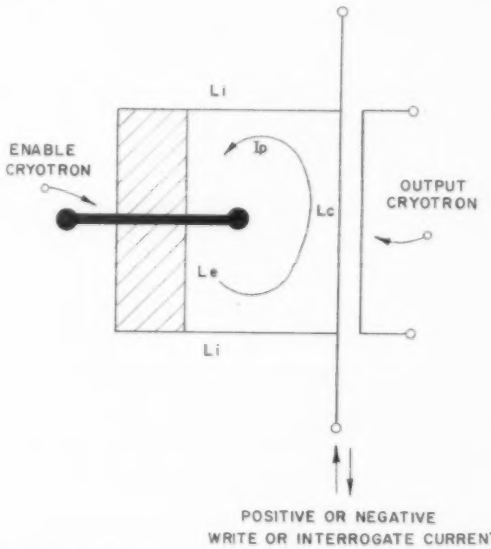


FIG. 2. Circuit diagram of persistent-current memory cell.

which is used in the write cycle and an output cryotron which is used in the read cycle. The control and gate of the output cryotron are of equal widths and parallel to each other. Although the output gate is shown beside its control in the figure, it is placed directly under the control in the actual circuit. The parallel gate and control configuration

is possible because it is not necessary for this cryotron to have high current gain. The advantage of this arrangement is that resistance is restored to the complete length of the output gate, a relatively high resistance is obtained, and the time constant of the output-gate network is reduced.

To establish a persistent current in the cell the control of the enable cryotron is energized causing the enable gate to become resistive; then the write current is turned on. All of the write current I_w will flow through the superconducting path of the output control. When the enable control current is switched off and the enable gate becomes superconducting again, the write current will continue to flow in the output control.

The magnetic flux $L_c I_w$ is now trapped inside the superconductive loop and will remain constant as long as the enable gate remains superconducting. If the write current is turned off a circulating current I_p will be established in the loop so as to maintain a constant flux through the loop. The value of persistent current is determined by equating the product of inductance and currents before and after the write current is turned off.

$$L_c I_w = (L_c + 2L_1 + L_e) I_p$$

or

$$I_p = \frac{L_c I_w}{(L_c + 2L_1 + L_e)} \tag{4}$$

where L_c is the inductance of the output control,

L_1, is the inductance of one side of the loop,

L_e is the inductance of the enable gate.

When the memory is interrogated, the interrogate current I_t divides between the two paths of the memory cell according to the inverse ratio of their inductances. If these inductances are chosen properly then the sum of the persistent current and the interrogate current in the output control will be greater than a critical value $(I_c)_{\text{crit}}$ and the difference of these two currents will be zero. When $L_c = L_e + 2L_1$, this relationship will be obtained.

This choice of inductances will cause the interrogate current to divide equally between the output control and the enable gate. If the interrogate current is equal to $(I_c)_{\text{crit}}$ and the persistent current is $(I_c)_{\text{crit}}/2$ then the current in the output control will be either $(I_c)_{\text{crit}}$ or zero.

Substituting $L_c = L_e + 2L_1$ into equation (4) we find that $I_p = \frac{1}{2} I_w$. Therefore the persistent current will be equal to $(I_c)_{\text{crit}}/2$ when $I_w = (I_c)_{\text{crit}}$.

For some circuits it may be desirable to produce a critical current in the output control regardless of the direction of the persistent current. The magnitude of this override current I_0 when it is opposing the persistent current, is given by

$$(I_c)_{\text{crit}} = \frac{I_0}{2} - I_p \tag{5}$$

therefore $I_0 = 2[(I_c)_{\text{crit}} + I_p]$ and since $I_p = (I_c)_{\text{crit}}/2$ then $I_0 = 3(I_c)_{\text{crit}}$.

THE ENABLE CRYOTRON

So far, the dimensions of the enable cryotron have not been discussed. It should be noted that when the override current is opposing the persistent current in the output control it is, at the same time, adding to the persistent current in the enable gate. The sum of the currents in the enable gate is equal to $2(I_c)_{\text{crit}}$.

If the memory cell is designed so that the override current can be used then the enable gate must be capable of carrying $2(I_c)_{\text{crit}}$ without self-quenching. The enable gate must therefore be 20–30 times wider than the output control[9]. If the override feature is not used the gate need be only 10–15 times wider than the output control.

The enable control is driven by an external source and the amount of drive current required is directly proportional to the width of the control. A wide enable control produces a large resistance in the enable gate and thus a short write time; therefore, the enable control should be made as wide as possible considering the capabilities of the external drive source.

To summarize: (1) The memory cell should be constructed so that the inductance of the cell is twice the inductance of the output control. (2) I_w should equal $(I_c)_{\text{crit}}$. (3) I_p will then equal $(I_c)_{\text{crit}}/2$. (4) Interrogation with $+(I_c)_{\text{crit}}$ will produce a current equal to $(I_c)_{\text{crit}}$ in the output control. (5) Interrogation with $-(I_c)_{\text{crit}}$ will result in approximately zero current in the output control. (6) An override current of $3(I_c)_{\text{crit}}$ will produce a greater-than-critical current in the output control regardless of the direction of the persistent current. (7) The enable gate should be 20–30 times wider than the output control.

TEST RESULTS

A number of memory cells have been constructed and tested. Fig. 3 is a photograph of one of these cells. The parameters that are of interest in testing the cell are listed below:

$(I_c)_{crit}$ = critical controlling current, determined at a specific gate current (for this measurement the enable gate must be resistive)
I_g = current in the output gate
$(I_g)_{crit}$ = critical output-gate current, determined in the absence of control current
I_w = write current
I_i = interrogate current
I_p = persistent current in cell
$(I_{eg})_{crit}$ = critical current of the enable gate
$(I_p)_{crit}$ = maximum persistent current that can be stored in the memory cell

All the above parameters can be measured directly, except I_p, $(I_p)_{crit}$ and $(I_{eg})_{crit}$. I_p can be calculated from formula (4). The maximum persistent current that can be stored in the cell is usually limited by the critical current of the enable gate and therefore $(I_{eg})_{crit} = (I_p)_{crit}$. This current can be determined by attempting to store a large persistent current and allowing this current to decay. The current that remains can be determined by measuring the magnitude of I_i under these conditions.

$$(I_p)_{crit} = (I_{eg})_{crit} = (I_c)_{crit} \pm \frac{I_i}{2} \qquad (6)$$

[the sign is minus if $(I_p)_{crit} < (I_c)_{crit}$ and plus if $(I_p)_{crit} > (I_c)_{crit}$.]

Several memory cells have been tested and were found to operate satisfactorily. The measured current values for a typical cell are as follows:

Table 1

Cell no. R1	
Temperature	= 3·60°K
$(I_g)_{crit}$	= 0·150 A
$(I_c)_{crit}$	= 0·62 A (with I_g = 0·025 A)
$(I_{eg})_{crit} = (I_p)_{crit}$	= 1·2 A (with I_g = 0·025 A)
I_w	= 0·6 A
I_i	= 0·64 A

One of the memory cells was tested to determine its response to pulses. A circulating current was loaded into the cell, and it was interrogated in both the positive and negative directions with a 1-μsec pulse from a Dumont 404 pulse generator. Figs. 4–6 are oscillograms of the pulse output from the cell. Fig. 4 is the memory output during a positive interrogation when there is no circulating current in the cell; Fig. 5 is the output during a positive interrogation with a positive circulating current in the cell; and Fig. 6 is the output during a negative interrogation with a positive circulating current in the cell.

The output signal of a "yes" comparison is clearly distinguishable from the output signal of a "no" comparison. The memory cell can undoubtedly be interrogated in much less than 1 μsec but in this test the rise time was limited by the type 53/54D Tektronix oscilloscope preamplifier. In future tests it may be possible to use a step-up transformer on the output of the cell so that a lower-gain preamplifier can be used.

TIME CONSTANT OF CATALOG MEMORY

Interconnection of the memory cells in the catalog memory is shown in Fig. 7. The enable leads are arranged so that a single enable lead will make all of the cells in one word resistive. To write a new word into the memory the proper enable lead is energized and write currents are applied on the write–interrogate leads. The same inputs are used for both write and interrogate.

The time required for an interrogation is determined by the time constant associated with the redistribution of the output-gate current within the gate network. For each word in the memory there are n parallel gate paths in which there may be one or more superconducting paths, depending on the stored word and the interrogation word. The most unfavorable situation occurs when current initially distributed evenly among n resistive gates must be redistributed into a single superconducting gate. Quenching patterns with more than one superconducting path will have shorter time constants.

The connection of additional words in series in the memory does not alter the time constant for redistribution in each word, but the transient voltage appearing across each word will be additive. In general, the time required for an interrogation

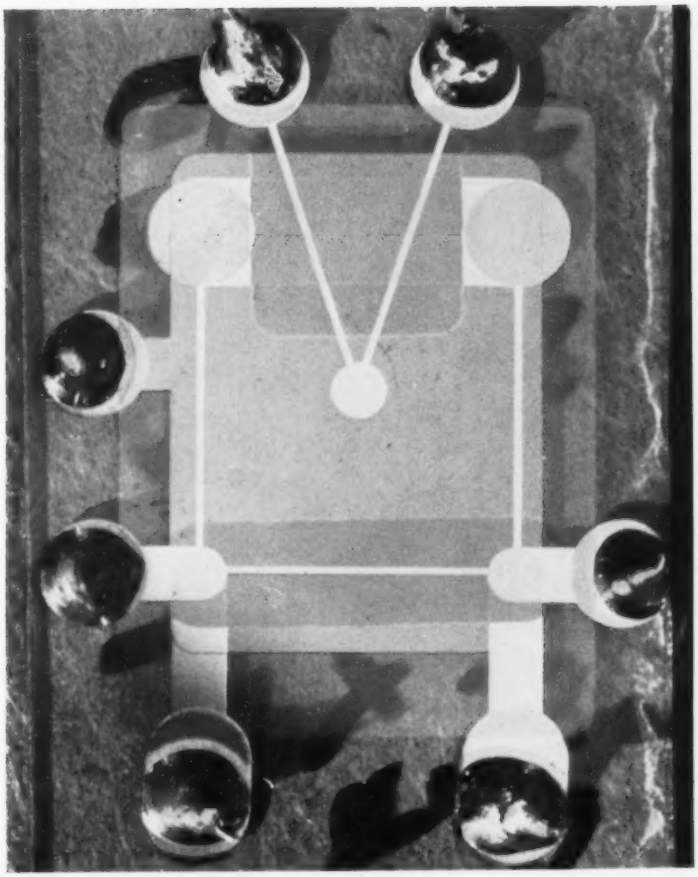


FIG. 3. Persistent-current memory cell.

[*facing p.* 360

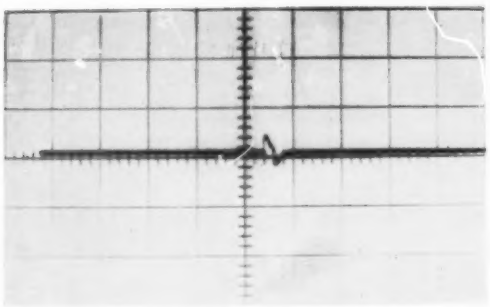


FIG. 4. Memory-cell output when $I_p = 0$ and I_i is positive (y-axis = 1 mV/cm; x-axis = 5 μsec/cm).

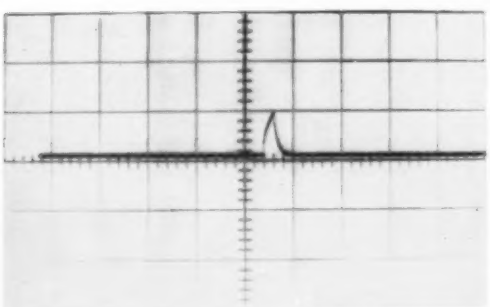


FIG. 5. Memory-cell output when I_p is positive and I_i is positive (y-axis = 1 mV/cm; x-axis = 5 μsec/cm).

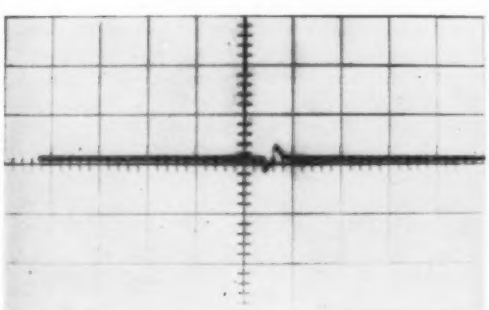


FIG. 6. Memory-cell output when I_p is positive and I_i is negative (y-axis = 1 mV/cm; x-axis = 5 μsec/cm).

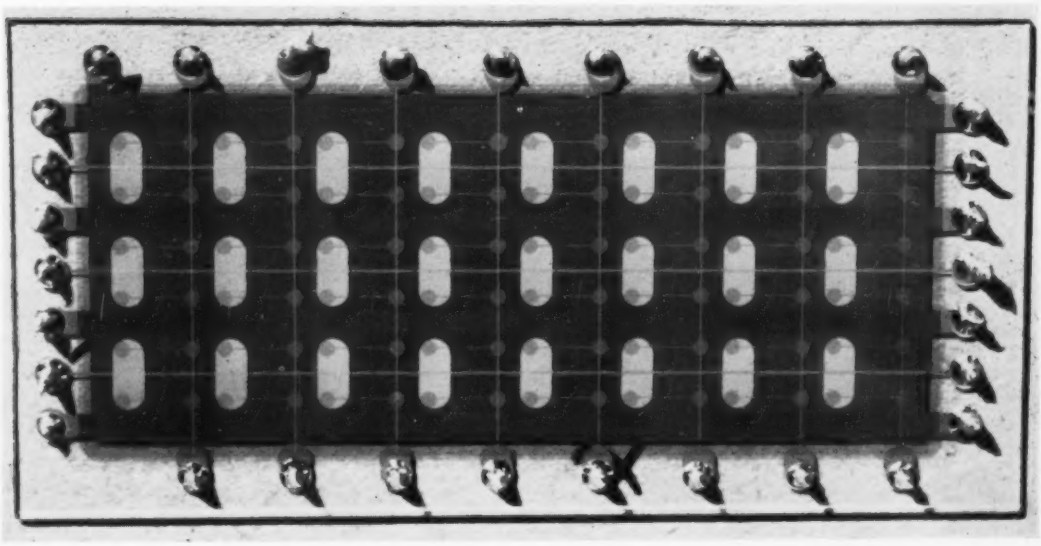


FIG. 8. Photograph of thin-film catalog memory containing 3 eight-bit words.

BLANK
PAGE

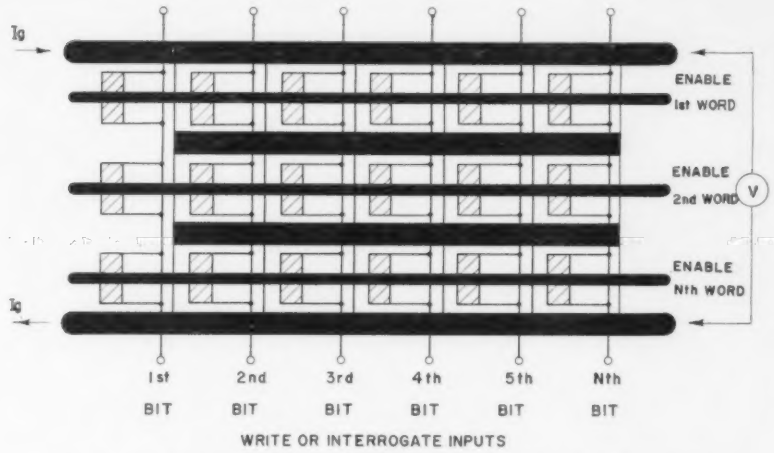


FIG. 7. Circuit diagram of catalog memory containing 3 six-bit words.

is given by:[10]

$$T = n \frac{L_g}{R_g} \ln N \qquad (7)$$

where n is the number of bits per word, L_g/R_g is the time constant of a low-gain cryotron and N is the number of words.

In a previous section the time constant for a low-gain cryotron was calculated to be of the order of 10^{-11} sec (regardless of its length or width). In this circuit, however, the gate current will be divided between a number of paths and will therefore be small. Unfortunately, for small gate currents considerably more control current is required to produce fully restored gate resistance. If the restored resistance is $R_g/10$ then L_g/R_g would be 10^{-10} sec. On this basis the interrogation time for a memory containing 4000 twenty-five-bit words would be approximtaely 2×10^{-8} sec.

Fig. 8 is a photograph of a memory containing three eight-bit words. Several memories of this type have been constructed and are now being tested. No attempt at miniaturization has been made with the present memories; however, an equivalent parts density of 10^6 bits per cubic foot should be feasible with presently known techniques.

A PROPOSED STORAGE AND RETRIEVAL SYSTEM

One of the most interesting uses of a catalog memory is in a storage and retrieval system. A number of catalog memories can be interconnected as shown in Fig. 9. Each word of the catalog memory is made longer and is divided into several sections. The example shown in Fig. 9 is a system for cataloging library information.

One or more sections of the memory can be interrogated at the same time in the same manner that a single catalog memory is interrogated. In order to obtain a yes or no output, however, it will be necessary to make resistive each of the sections that is not being interrogated. This is accomplished with the override current discussed in a previous section.

For example, if the memory is interrogated with the name of the title and author, then the override current is applied to the location section. When a "yes" answer is obtained, the word stored in the location section (corresponding to that title and author) is read out as follows:

The output is obtained in a serial manner by momentarily reducing the interrogate current in each bit from I_0 to $+I_t$. This is the same as interrogating each bit, individually, with a positive interrogate current, and therefore a "yes" output will indicate that a positive circulating current is stored in that bit and a "no" output will indicate that a negative circulating current is stored in that bit. A record of the "yes" and "no" bits obtained in the serial readout process will reveal the location of the publication.

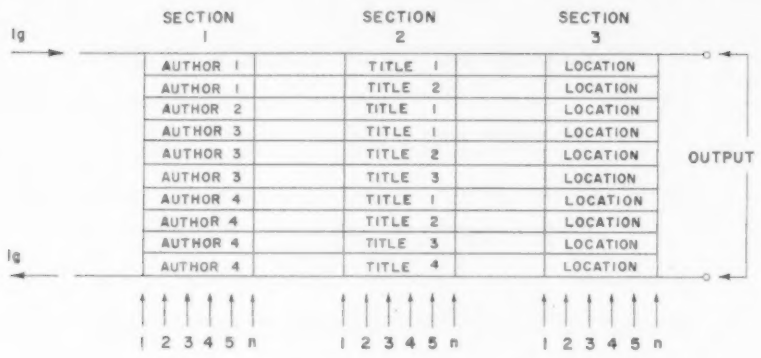

SECTION 1
SECTION 2
SECTION 3
Ig
Ig
AUTHOR 1 TITLE 1 LOCATION
AUTHOR 1 TITLE 2 LOCATION
AUTHOR 2 TITLE 1 LOCATION
AUTHOR 3 TITLE 1 LOCATION
AUTHOR 3 TITLE 2 LOCATION
AUTHOR 3 TITLE 3 LOCATION
AUTHOR 4 TITLE 1 LOCATION
AUTHOR 4 TITLE 2 LOCATION
AUTHOR 4 TITLE 3 LOCATION
AUTHOR 4 TITLE 4 LOCATION
OUTPUT
1 2 3 4 5 n
1 2 3 4 5 n
1 2 3 4 5 n
WRITE, INTERROGATE OR OVERRIDE INPUTS

FIG. 9. Use of catalog memories in a storage and retrieval system.

CONCLUSION

Considerable research effort has been expended on the thin-film cryotron and, at present, the device appears to have a promising future. If cryotron circuits are to gain acceptance in the computer field they must be able to solve problems either faster or more economically than is possible with competitive devices.

One of the first applications where the cryotron will find acceptance may be in a system combining transistors and cryotrons. The storage and retrieval system outlined in this paper is an excellent example of the use of cryotrons to solve a problem which is difficult and expensive to solve with other devices.

REFERENCES

1. C. R. SMALLMAN, A. E. SLADE and M. L. COHEN, *Proc. I.R.E.* **48,** 1562 (1960).
2. V. L. NEWHOUSE, J. W. BREMER and H. H. EDWARDS, The crossed film cryotron and its application to digital computers. *Proceedings of the Eastern Joint Computer Conference, December,* 1959, p.255.
3. C. R. SMALLMAN, An evaporated film cryotron circuit. Presented at the *Solid-State Circuits Conference, February,* 1960.
4. E. C. CRITTENDEN, JR., A computer memory element employing superconducting persistent currents. Presented at the *Fifth International Conference on Low Temperature Physics and Chemistry, August,* 1957.
5. M. J. BUCKINGHAM, The persistatron: a superconducting memory and switching element for computers. Presented at the *Fifth International Conference on Low Temperature Physics and Chemistry, August,* 1957.
6. J. W. CROWE, *IBM J. Res. Dev.* **1,** 295 (1957).
7. L. L. BURNS, G. W. LECK, G. A. ALPHONSE and R. A. KATZ, *Solid-State Electron.* **1,** 343 (1960).
8. A. E. SLADE, H. O. MCMAHON, A cryotron catalog memory system. *Proceedings of the Eastern Joint Computer Conference, December,* 1956, p. 115.
9. A. E. SLADE, *Proc. I.R.E.* **48,** 1569 (1960).
10. E. C. PARK, *Cryotron Memory Systems.* Final Progress Report, Signal Corps Contract No. DA49–170–sc–1986, Arthur D. Little, Inc. (1957).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 363–367. Printed in Great Britain

CHARACTERISTICS OF BULK AND THIN-FILM SUPERCONDUCTING ALLOYS

A. M. TOXEN

Research Laboratory, International Business Machines Corporation, Poughkeepsie, New York

INTRODUCTION

THE ultimate limitations of a superconducting computer circuit are determined by the fundamental properties of the materials from which the circuit is constructed. If one considers a thin-film cryotron loop, the speed with which currents can be switched will be approximately proportional to the normal state resistance of the cryotron gate. This will, in turn, be proportional to the intrinsic resistivity of the gate material. Since the gate is switched from superconducting to resistive by the magnetic field of a control current, it is desirable that the gate critical field be conveniently low.

To compare the properties of films varying in composition and thickness, it is useful to define a figure of merit, i.e. a measure of the suitability of a particular film for use as a cryotron gate. One which has been implied is ρ_0/h_c, the ratio of residual resistivity to film critical field. By definition, the gain of a cryotron is I_g/I_c, the ratio of the gate critical current to the control current necessary to switch the gate. Clearly the control current I_c is proportional to the gate critical field h_c. For films which are thicker than a penetration depth, the gate critical-current density, at a given temperature and for a given film width, is proportional to the critical field of a bulk specimen that has the same composition as the film.[1,2] Since the bulk critical field at a given temperature is nearly independent of composition for a dilute alloy system, the gain-resistance product for the films to be discussed is, on the basis of these simple considerations, proportional to ρ_0/h_c. This then can be considered a figure of merit and, accordingly, will be used in the discussion to follow.

BULK ALLOYS

To increase the electrical resistivity of a metal one must introduce imperfections such as impurities, dislocations or grain boundaries. However, alloys, compounds, and strained materials frequently exhibit properties quite different from pure unstrained monatomic materials,[3–6] e.g. high critical fields, incomplete Meissner effect, broadness and non-coincidence of magnetic and resistance transitions, and sensitivity of resistance transition to measuring current. On the other hand not all alloys are found to exhibit the above properties. Stoichiometric compounds and alloys prepared with care have properties similar to the ideal ones of pure elemental superconductors[7–9] suggesting that the non-ideal properties may result from microscopic inhomogeneities. To investigate this latter conjecture, REEBER undertook a careful study of the indium–mercury alloy system in the composition range of 0–7 atomic per cent Hg.[10] The specimens were prepared as follows:

(1) After weighing, the materials were melted together in vacuum and quenched to avoid long-range segregation.
(2) The resulting alloy was then extruded through a 1-mm die to obtain further mechanical mixing.
(3) Finally, the extruded rods were annealed at 90 per cent of the solidus temperature.

Some of REEBER's results are summarized in Figs. 1–7. In Fig. 1 the residual resistivity ratio (the ratio of the resistivity at 4·2°K to the room-temperature resistivity minus the resistivity at 4·2°K) is plotted as a function of mercury content. This quantity varies linearly with mercury content from 10^{-4} for pure indium to 0·074 for In + 5 atomic per cent Hg. A typical electrical resistance transition for an alloy specimen is shown in Fig. 2. The ratio of measured resistance to normal state resistance

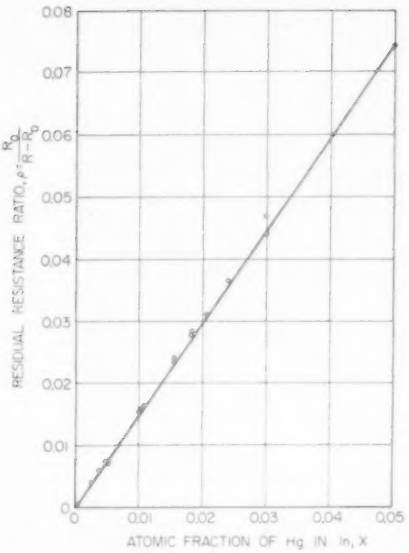


FIG. 1. The residual resistance ratio of In–Hg alloys.

is plotted versus magnetic field. This plot shows the distinction between the "sharp" portion of the transition which coincides with the magnetic transition and the broad portion which varies with the thermal history of the specimen as well

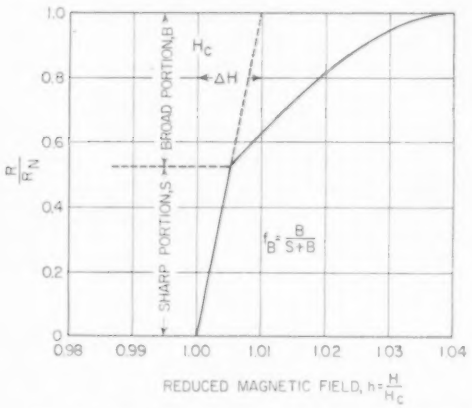


FIG. 2. A typical resistance transition of an alloy.

as the current density in the specimen. The manner in which the broad portion of the transition can be reduced by increasing the current density is demonstrated in Fig. 3. Fig. 4 shows the important result that the broad portion of the transition can be considerably reduced through proper annealing. The variation of T_c, the superconducting transition temperature, with composition is shown in Fig. 5. The existence of the minimum

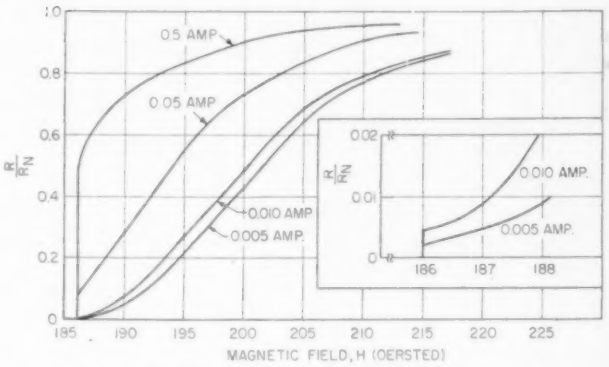


FIG. 3. The effect of measuring current on the resistance transition. The sample contains 6·6 atomic per cent Hg and was annealed at 296°K for 100 hr.

is quite typical of most alloy systems.[11,12] The variation with composition of the width of the sharp portion of the transition is indicated in Fig. 6. It is very important to note that the minimum transition width occurs not for pure indium, but at the composition for which T_c vs. composition exhibits a minimum ($\partial T_c / \partial x = 0$). Hence it may be concluded that one of the factors influencing

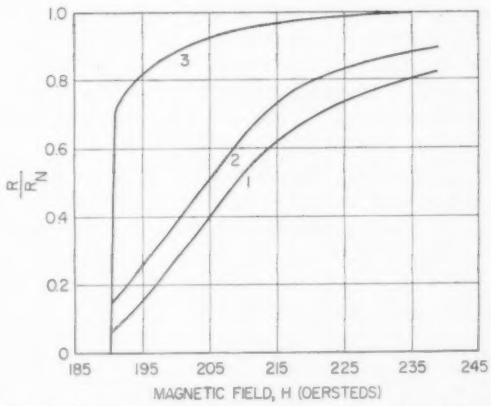


FIG. 4. Resistance transitions for specimens with different thermal histories. Composition is In+4·98 atomic per cent Hg; $t = 0{\cdot}62$; measuring current equal to 0·100 A. Curve 1—296°K for 725 hr; curve 2—368°K for 100 hr; curve 3—383°K for 100 hr.

transition width is the variation of imperfection concentration throughout a specimen, *even in a very pure specimen*. In Fig. 7 is plotted the resistance transition of a specimen containing 1·85

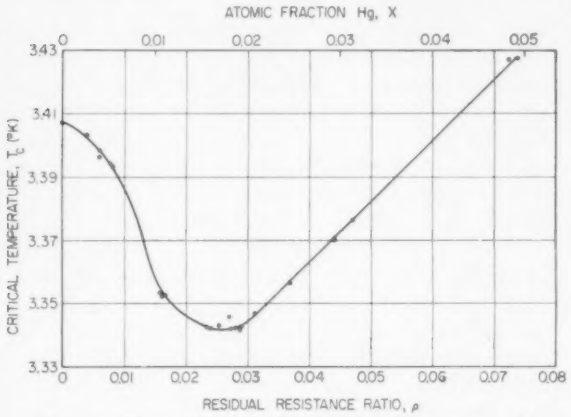


FIG. 5. The variation of critical temperature with mercury content.

atomic per cent Hg, a composition occurring near the minimum of the T_c vs. composition curve (see Fig. 6). Clearly, this transition, whose width is less than 0·001 per cent of T_c, is extremely sharp.

Finally, REEBER's measurements indicate that, for the composition range described here (0 7 atomic per cent Hg), the specimen critical fields at a given reduced temperature, T/T_c, differ from that of pure indium by no more than a few per cent.

In summary, by alloying, one can produce materials whose superconducting transitions are as "ideal" as those of the best elemental specimens, but whose residual resistivities are a thousandfold larger.

THIN-FILM ALLOYS

Because of the high residual resistivities obtainable with alloys, the study of thin alloy films was begun. The first system chosen for examination was indium and tin. This system was chosen because of its similarity to indium and mercury and because of the ease of thin-film fabrication—those in the present study have been produced by evaporation from a single indium and tin source.

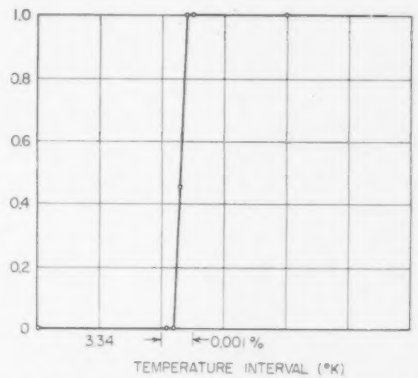


FIG. 7. The transition to superconductivity of indium with 1·85 atomic per cent Hg at $H = 0$, measuring current equal to 0·010 A.

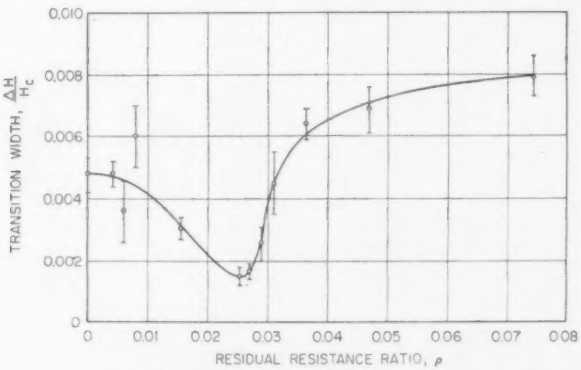


FIG. 6. The width of the resistance transitions for dilute In–Hg alloys.

Films have been evaporated onto fused-quartz substrates held at room temperature and at 77°K. The films vary in composition from 0 per cent Sn to 3 atomic per cent Sn. The thicknesses so far studied are in the range of 3000–9000 Å for the alloy films, and in the range of 2500–100,000 Å for the indium films.

The film properties which were measured directly were: the physical dimensions; the electrical resistance at room temperature, 77°K and 4·2°K; and the variation of the critical field with temperature as determined by resistance transition measurements. To calculate the penetration depth from the critical-field measurements, one must use some theoretical relationship between the two. The simplest relationship is that given by the LONDON theory[13] which is the following:

$$\frac{H_c^2}{h_c^2} = \frac{1-[\lambda(t)/a]\tanh[a/\lambda(t)]}{1+(\alpha_n-\sigma_s)8\pi/aH_c^2} \tag{1}$$

where h_c is the film critical field; H_c is the critical field of a bulk specimen composed of the same material as the film; $\lambda(t)$ is the penetration depth, which varies with reduced temperature; $t \equiv T/T_c$; the quantity a is the film half-thickness; and $\alpha_n-\alpha_s$ is the difference in surface energies between the normal and superconducting states. Ordinarily $8\pi\ (\alpha_n-\alpha_s)/aH_c^2$ is assumed to be small compared to 1, and equation (1) is simplified to

$$H_c^2/h_c^2 = 1-\frac{\lambda}{a}\tanh\frac{a}{\lambda} \tag{2}$$

The quantity λ is derived by the LONDON theory to be

$$\lambda_L = \left(\frac{mc^2}{4\pi n_s e^2}\right)^{1/2} \tag{3}$$

In equation (3), c is the velocity of light, m the electronic mass, e the electronic charge and n_s is the density of superconducting electrons. If the temperature variation of n_s is taken to be that given by the GORTER–CASIMIR two-fluid model[14], one obtains the familiar temperature dependence of the penetration depth which is found to be obeyed to a high degree of accuracy by a number of superconducting materials[15]

$$\lambda = \lambda_0(1-t^4)^{-1/2} \tag{4}$$

In equation (4) λ_0 is the value of λ at $t = 0$.

For the indium and indium–tin films measured so far, the critical-field data fit reasonably well the relation given by equation (2) if one assumes the temperature dependence of the penetration depth to be that given by equation (4). LOCK[16], from magnetic measurements, obtained similar results for thin indium films.

Using equation (2) effective penetration depths can be calculated, and these will be published in the near future.

In Table 1 are shown some of the preliminary results. The critical field chosen for reference purposes is the measured-film critical field at $t = T/T_c = 0{\cdot}6$. The quantity $\rho_0/h_c\ (t = 0{\cdot}6)$ is considerably larger for the alloy films than for the pure indium films and seems to increase with increasing film thickness and tin content. It therefore seems reasonable to conclude that alloy films deserve serious consideration for use in superconducting thin-film computer circuits.

Table 1

Film thickness (Å)	Composition (At. % Sn)	1000 $\rho_0 h_c$ $(t = 0{\cdot}6)$ ($\mu\,\Omega$-cm/G)
110,000	Pure In	0·044
4700	Pure In	0·452
6000	Pure In	0·505
2400	Pure In	0·633
2400	Pure In	0·820
2500	Pure In	0·904
2600	Pure In	0·802
5000	1·4	1·72
3200	2·8	3·57
2730	1·5	1·96
3800	1·5	1·98
3400	1·8	2·38
3400	1·8	2·30
9000	1·6	2·49
9000	1·6	2·44

Acknowledgements—The author wishes to express his appreciation to W. B. ITTNER, III for many helpful discussions; to M. D. REEBER for the use of his data and Figs. 1–7; and to W. F. HEYDMAN, R. E. HORSTMANN and R. J. PARKER for the thin-film measurements presented.

REFERENCES

1. W. B. ITTNER, III, *Proceedings of the Symposium on Superconductive Techniques for Computing Systems,*

Washington, 1960, p. 239. Office of Naval Research (1960).
2. K. T. Rogers. Thesis, University of Illinois (1960).
3. D. Schoenberg, *Superconductivity*. Cambridge University Press (1952).
4. W. J. de Haas and J. Voogd, *Commun. Kamerlingh Onnes Lab. Leiden*, No. 208b (1930).
5. P. R. Doidge, *Proc. Phys. Soc. Lond.* **248**A, 553 (1955).
6. B. G. Laserew and A. A. Galkin, *J. Phys., Moscow* **8,** 371 (1944).
7. D. Schoenberg, *Nature* **142,** 874 (1938).
8. B. G. Laserew and I. E. Nakhutin, *J. Phys., Moscow* **6,** 116 (1942).
9. J. W. Stout and L. Guttman, *Phys. Rev.* **79,** 396 (1950).
10. M. D. Reeber, *Phys. Rev.* **117,** 1476 (1960).
11. E. A. Lynton, B. Serin and M. Zucker, *J. Phys. Chem. Solids* **3,** 165 (1957).
12. G. Chanin, E. A. Lynton and B. Serin, *Phys. Rev.* **114,** 719 (1959).
13. F. London, *Superfluids* Vol. 1, p. 130. Wiley, New York (1950).
14. C. J. Gorter and H. G. B. Casimir, *Physica, 's Grav.* **1,** 306 (1934).
15. B. Serin, *Encyclopedia of Physics* Vol. IV. Springer, Berlin (1956).
16. J. M. Lock, *Proc. Roy. Soc.* A**208,** 391 (1951).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 368–377. Printed in Great Britain

EFFECT OF DEFECTS ON THE SUPERCONDUCTING PROPERTIES OF TANTALUM

D. P. SERAPHIM

Research Laboratory, International Business Machines Corporation, Poughkeepsie, New York

INTRODUCTION

MANY prospective cryogenic devices utilize the superconducting property of metals and the transition to a resistive normal state brought on by a magnetic field. Certain characteristics, such as the electrical resistivity, the thermal resistivity, the reversibility of the transition, the magnetic-field interval which the transition covers, the critical magnetic field and critical temperature, and the reproducibility of the latter characteristics, are important in the operation of these cryogenic devices. These characteristics are often neither ideal nor reproducible and it is frequently difficult to assign responsibility to any specific defect either in the metallic superconductor or in the building of the device. There is no doubt, however, that all types of lattice defects have an influence on the characteristics of the resistance transition. Indeed the residual resistivity, itself, is a quantitative measure of the density of lattice imperfections in any metal, unless that metal is so free from imperfections that the mean free path approaches the spacing between the bounding surfaces.

Bulk specimens of metals as opposed to thin films are more suitable for investigation of the effects of specific lattice defects, for such specimens are not complicated by magnetic-field penetration effects in the superconducting state, nor by the constraints imposed by substrates. Moreover, the non-ideal behavior of a variety of materials appears to be independent of geometry this being particularly true with respect to the nature of the return of normal-state resistivity in the presence of a magnetic field. Thus, while the present experiments are on bulk tantalum, the results can be expected to be generally applicable.

In an earlier paper[1] it was shown that the Meissner–Ochsenfeld effect often occurs at fields significantly lower than the field which returns full normal-state resistivity. This was long ago observed by MENDELSSOHN[2] in lead and later investigated in some detail by DOIDGE[3] for tin containing indium as an impurity. More recently REEBER[4] has found this effect in indium, with mercury as an impurity, while MAPOTHER *et al.*[5] have investigated the irreversibility of transitions in impure and deformed lead. Preliminary experiments with tantalum[1,6] have shown how sensitive the resistance transition can be to the presence of specific interstitial impurities.

An interpretation[7] of the observed phenomena, based on the influence of the mean free path in determining a negative surface energy in the interfaces between superconducting and normal regions, has had qualitative success in explaining the simultaneous existence of the two states at a single magnetic field and temperature. A negative surface-energy term in the free energy allows the phases to subdivide without size restriction, with the consequence that field penetration into the superconducting domains becomes important in determining the susceptibility. The magnetic energy of the superconducting regions is appropriately reduced and high fields become a necessity for producing a transition between states. The presence of a reasonable Meissner–Ochsenfeld effect (i.e. almost 100 per cent flux exclusion in the superconducting state)[8], however, is in serious contradiction to such a theory.

To circumvent the contradiction, the negative surface energy must be localized and isolated to small regions (a substructure within the specimen), and it may be inferred that these regions have

different physical properties from the bulk of the metal. Determination of the superconducting properties of the substructure can thus lead to elucidation of its physical nature. This approach has been supplemented in the present work with X-ray diffraction and electron transmission microscopy.

PREPARATION OF SPECIMENS

Variations between various reported superconducting properties of the transition metals (see the critical-field curves of WHITE *et al.* and others[9]) have often been caused by their different vacuum and thermal pretreatments, which have a profound effect on many properties of these exothermic occluders of gaseous impurities. In our work, considerable effort was devoted to purification[10] of the tantalum 0·010-in. diameter wire specimens by heating them to near the melting point in ultra-high vacuum ($\sim 10^{-9}$ mm Hg). The residual resistivity (ρ_i) then approaches $1 \times 10^{-3}\ \mu\Omega$ cm* and the transitions become ideal in the sense that susceptibility and resistivity transitions occur at the same temperature and magnetic field, and flux trapping is negligible. The critical temperature increases with increasing purity (decreasing ρ_i) to the highest value so far reported for tantalum[11,12], 4·483 ± 0·003°K.

The system, employed for absorption of interstitials from the gas phase into the metal, incorporates, in series sequence: fore pump, oil diffusion pump, liquid-nitrogen cold trap, Veeco ion gauge and pump, flask containing the specimens, Veeco variable leak valve, and flask containing the gas. Kovar graded glass seals are used for electrical leads into the system enabling the entire system to be baked above 400°C.

After the tantalum was purified and the pressure in the system dropped to an ultimate as low as 10^{-9} mm Hg, time was allowed for reasonable diffusion and evaporation of substitutional impurities. The leak valve was then opened until a desired pressure was registered by the gauge. The system was held at a fixed pressure and temperature for times much longer than required to obtain an equilibrium between the gas phase and the heated tantalum (see Ref. 10 for appropriate time constants). The temperatures and pressures employed were varied according to the concentration of gas desired in the specimens, the higher pressures and lower temperatures providing the greater concentration.

The resistivity at helium temperatures (ρ_i) increases by $\sim 5{\cdot}1\ \mu\ \Omega$-cm for each atomic per cent of either nitrogen or oxygen[13,14] in tantalum, while the resistivity from phonon scattering alone, ρ_T, is $13{\cdot}6\ \mu\ \Omega$-cm at 298°K[10] and is negligible at 4·2°K. Hence calculation of the atomic concentration c in per cent is accomplished for these interstitials by determining the parameters in the following equations

$$c = \frac{13{\cdot}6}{5{\cdot}1}\,\frac{R_{He}}{R_T - R_{He}} \equiv \frac{13{\cdot}6}{5{\cdot}1}\,\frac{1}{\Gamma}$$

where R_T and R_{He} are the resistances of the specimen at ambient and at liquid-helium temperatures respectively. Correction can be made for the effect of residual impurity prior to the absorption of the desired interstitial. The concentration will be reported here as a nominal concentration in terms of the ratio $1/\Gamma$, which is favored because of lack of precision in the conversion factors.

MEASUREMENTS

The 1958 temperature scale has been used throughout to determine the temperature from the absolute-pressure data taken with a U-tube manometer and Wilde cathetometer. The reproducibility of temperature is estimated as 3×10^{-3} °K. However, many of the measurements were made in reference to a standard specimen where the relative errors should be less than 10^{-3}°K.

Transitions of specimen resistance were measured and the critical field H_c, at which the bulk of the specimen switches from the superconducting to the normal state, was established by a procedure used previously.[1,12] The technique relies on the ability of an appropriately large measuring current† to quench the filamentary conduction near T_c. The transitions are then characterized by a sharp onset

* If the residual resistivity were due to scattering from interstitial impurities a reasonable estimate of the concentration would be $< 2 \times 10^{-5}$ per cent by weight.

† Although the current is sufficient to quench the filamentry conduction, the field it causes normal to the wire axis is insignificantly small, i.e. the Silsbee effect is negligible.

of resistance at H_c, as is shown in Fig. 1(a). The characteristics of the filamentary substructure can be observed by decreasing the measuring current until the potential between the leads monitoring 1 in. of the wire in the normal state is approximately ten times greater than the noise level. The

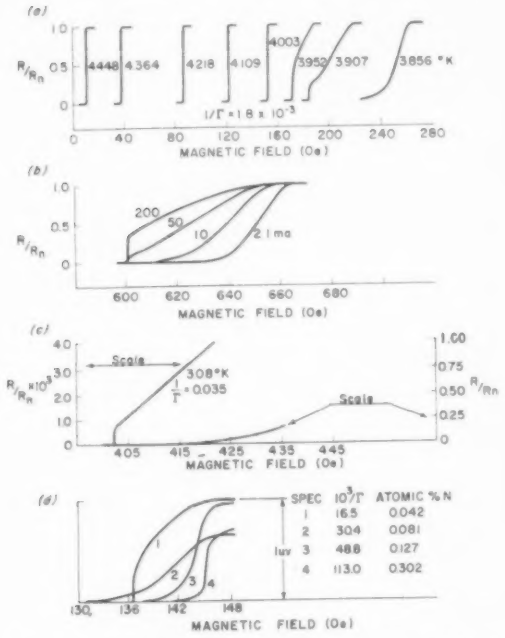


FIG. 1. Return of resistivity with magnetic field for tantalum with absorbed nitrogen: (a) at various temperatures, (b) with variable current density, (c) at high and low amplification and (d) with variable nitrogen content.

transitions are then characterized by a continuous increase of resistance with increasing magnetic field at fields higher than the field at which the sharp step may occur (see Fig. 1). With an integrator and a Leeds and Northrup amplifier the signal can be decreased to 1 or 2 μV.

Unfortunately the specimens with the least nitrogen must be measured at the highest current densities to keep the signal-to-noise ratio sufficiently high. The filamentary transition for the specimens with low nitrogen content are consequently broadened and shifted to fields closer to the bulk critical field. The technique is thus limited to specimens with a value of $\Gamma < 400$.

THE RESISTANCE TRANSITION AND THE EFFECT OF IMPURITIES

Resistance transitions typifying a variety of experimental conditions have been shown in previous investigations. The data will be reviewed here and complemented by data from recent work. Beginning with fixed low impurity content and current density the effect of varying the temperature is illustrated in Fig. 1(a). Near T_c, as the magnetic field is increased, the resistance transition is characterized by a sharp onset of resistance to the normal-state value. Independent but simultaneous measurements of the specimen susceptibility indicate that the bulk of the flux transition occurs at the same magnetic field.

As the temperature is decreased the sharp onset of resistance decreases in magnitude and is followed by a gradual restoration of resistance with increasing field. The sharp onset of resistance can be used to infer the position of the flux transition, while the gradual restoration of resistance with increasing field can be associated with a substructure representing, for specimens with dilute impurities, a small fraction of the total volume of the specimen. The fraction of resistance which is restored gradually and the detailed shape of this portion of the transition are a sensitive function of the measuring current and provide additional evidence that the superconducting inclusions are small for specimens with a dilute impurity content. The effect of measuring current on the shape of the transition is shown in Fig. 1(b). It is fairly evident that the smaller measuring currents create a favorable condition for observing a transition characteristic of the substructure and that the higher currents must be employed to observe the sharp onset of resistance. Susceptibility experiments were repeated under a variety of conditions and confirmed the fact that the sharp onset of resistance was coincident with an almost complete flux penetration.

As the nitrogen content (oxygen behaves similarly) is increased the transitions become less sensitive to changes in current density, and in practice it is difficult or impossible to create a sharp onset of resistance like the transitions of the form shown in Fig. 1(a) or (b). However, in the intermediate range of nitrogen content it is interesting to note that the substructure, although appearing to carry a continuous supercurrent, in

reality has a resistance that can be of the order of 10^{-4} of the normal-state resistance appearing at H_c. This behavior is illustrated by the data in Fig. 1(c).

In Fig. 1(d) where current density is kept at the lowest value, consistent with resolution from the amplifier, it is shown that the normal-state resistance returns at essentially the same magnetic field, independent of nitrogen content for a substantial range of concentrations. Thus an apparent critical field H_n can be established for the substructure. In summary H_c is the field producing full return of normal-state resistance at high current density. The amount of substructure is inferred qualitatively from the insensitivity of the position of the broad transition to measuring current density.

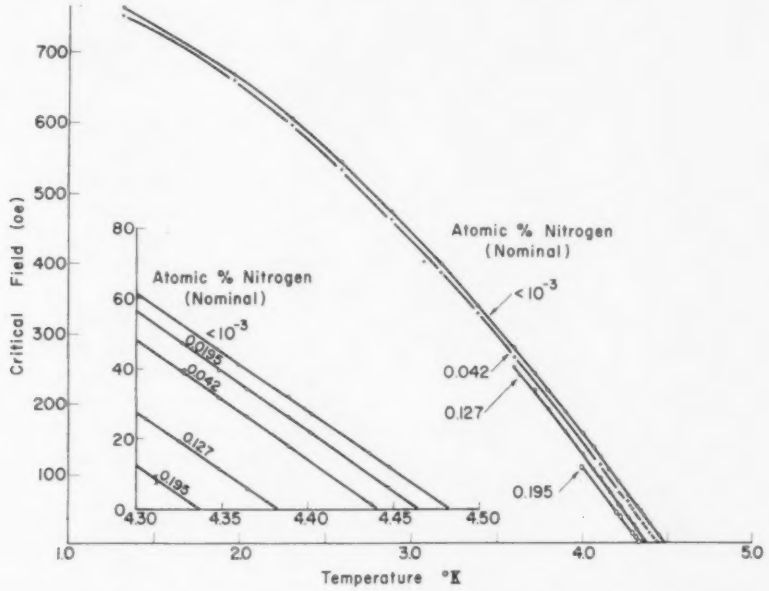


FIG. 2. Critical field curves for pure tantalum and for tantalum with absorbed nitrogen.

PROPERTIES OF THE BULK MATERIAL

For pure tantalum and for tantalum with a variety of nitrogen concentrations the critical-field curves are closely similar. The nitrogen interstitial in the tantalum lattice merely decreases H_c and T_c, as shown in Fig. 2, and within the precision of measurements the ratio (H_0^2/T_c^2)* remains constant. The region near T_c is expanded in the inset to illustrate similarity in another way, namely through the constancy of slope $\partial H_c/\partial T$ at $T = T_c$.

The change of critical temperature with interstitial nitrogen, oxygen and hydrogen and with cold work has been discussed in detail in a recent paper.[13] The magnitude of the effect for the oxygen, nitrogen or hydrogen interstitial solid solutions when correlated with residual resistivity, is in qualitative agreement with a mean-free-path effect and in quantitative agreement with the same effect in substitutional solid solutions as investigated by SERIN *et al.*[16,17] The decrease of T_c with increasing $1/\Gamma$ is shown for tantalum with interstitial nitrogen in Fig. 3. The results reported here have been extended to include solutions in the range of 2 atomic per cent nitrogen. The effect of cold work on T_c was also investigated but no significant change in T_c was found for the cold-

* Actually the similarity principle or the law of corresponding states as proposed by BARDEEN *et al.*[15] presents the ratio as $H_0^2/T_c^2 = K\gamma$ where γ is the temperature coefficient of electronic specific heat.

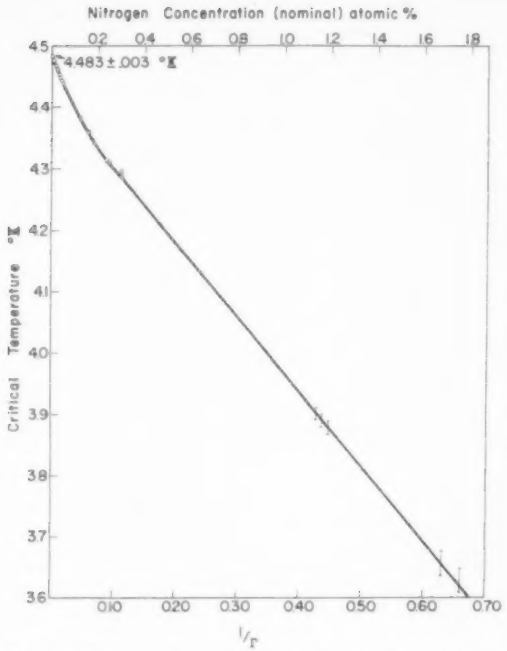


FIG. 3. The critical temperature for tantalum with absorbed nitrogen.

worked metal, although the mean free path was decreased by more than an order of magnitude.

PROPERTIES OF THE SUBSTRUCTURE

Interstitial impurities

The field H_n for return of the normal-state resistance is shown in Fig. 4 for specimens with nitrogen concentrations up to 0·25 atomic per cent. In the low-temperature region H_n is much greater than H_c, and extrapolation to 0°K yields an anomalously high value for H_0 in the order of 3000 oersteds. In the high-temperature range close to T_c, where $H_n < H_c$, the transition for the substructure disappears for specimens with $1/\Gamma < 0{\cdot}072$, and only the bulk transition is evident. However, as shown in Fig. 3, T_c and H_c decrease with increasing nitrogen content, thus with a gaseous concentration near 0·2 atomic per cent all transitions are broad in the immediate vicinity of T_c. In the range of composition greater than 0·2 atomic per cent the flux transitions are extremely broad. The magnetic field for the beginning, the end and the half point of the flux transition are plotted as a function of temperature for a specimen with 0·195 atomic per cent nitrogen in Fig. 5. Note that the critical field as defined

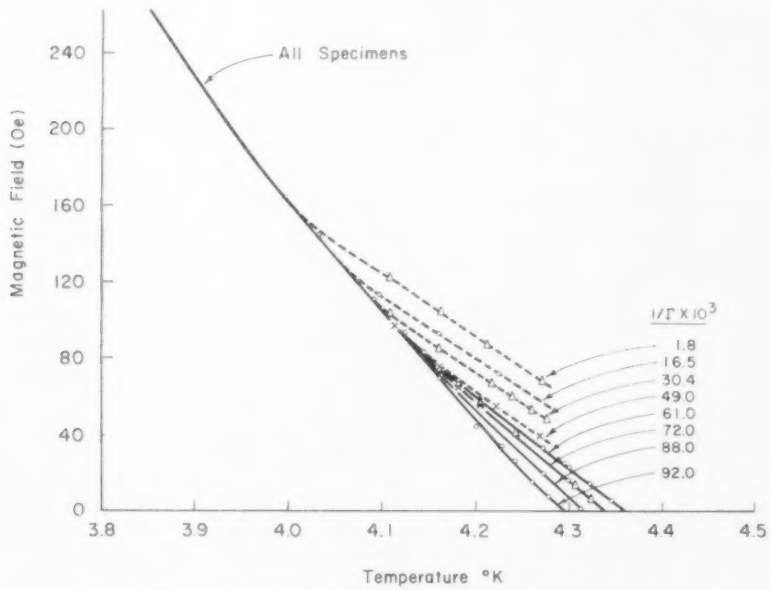


FIG. 4. The field curves for return of normal-state resistivity at low current density for tantalum with absorbed nitrogen.

from the half-value points obeys the similarity principle. The flux transition spreads to higher and to lower magnetic fields than the field predicted from calculations based on a slope of 300 oersteds/°K (the value for pure tantalum) and zero field at T_c. The insert in Fig. 5 shows

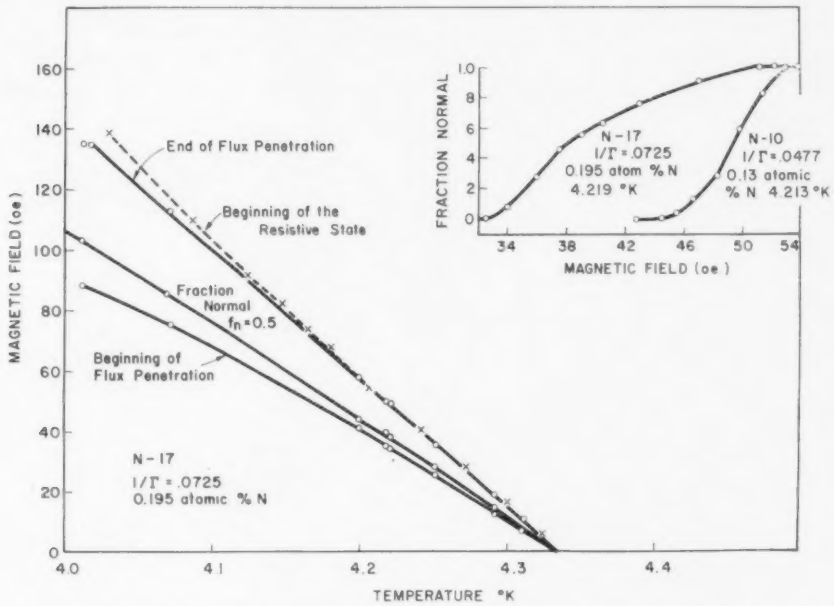


FIG. 5. Field penetration of highly impure tantalum and the field for initial appearance of resistivity as a function of temperature.

flux transitions at 4·2°K both for this specimen and for a specimen with considerably less nitrogen. Also shown in Fig. 5 is the field associated with the onset of the resistance transition. It may be observed that flux penetration is essentially complete when resistance appears. Even so, as previously mentioned, rather large current densities have little effect on the position of the resistance transition and it may be inferred that a substantial volume of the specimen is still superconducting when flux penetration is essentially complete. Therefore the size of the regions still superconducting must be in the range of the penetration depth or smaller to allow a complete field penetration and at the same time to remain superconducting. The fact that the break-up into an intermediate state depends not only on the nitrogen content but on the thermal history of the specimen in the low-temperature range (300°C, 20 hr) where nitrogen is just mobile, may be used to infer the presence of inhomogeneous distributions of nitrogen in the specimen.

X-ray investigation of the specimens and electron transmission micrography of thinned foils in the same composition range were initiated in an attempt to find a structure which might be responsible for the observed effects. One of the best techniques for obtaining Bragg reflections from a dilute Ta–N phase is by oscillation of single crystals in, for example, a Supper–Weisenberg type camera. With this method the diffracted beams are intensified into spots rather than spread into a Debye ring, as is the case with powder camera techniques. Even with the Supper–Weisenberg technique the reflections from the second phase were found to be extremely weak and in several specimens of appropriate composition could not be found at all. However when the reflections are present they take the positions shown in Fig. 6, half of which is a tracing of the

X-ray film and the other half a photograph. Between the zero- and first-order layer lines of the BCC lattice (rotated around ⟨110⟩) there are two extra layer lines obtained with a series of specimens containing from 0·01 to 1 atomic per cent nitrogen, the strength of the reflections increasing with nitrogen concentration. Extra reflections also appear near some of the primary BCC reflections, and with the better films two extra layers appear between the first- and second-order layers for the BCC lattice. These data can be interpreted as indicating the presence of two distinct and separate crystal phases in the same specimen. The primary crystal matrix is an interstitial solid solution of tantalum and nitrogen with a lattice parameter which increases with increasing nitrogen composition. The second phase is a superstructure, i.e. an ordered distribution of nitrogen which repeats every three BCC matrix cells. The lattice parameter of the superstructure is slightly larger than three times the spacing in the interstitial solid solution.

Electron transmission through thinned foils in the same range of composition also shows the presence of the second phase. The structure in Fig. 7 is a typical Widmanstätten pattern often seen in precipitating systems and is the result of selective growth of precipitate on certain matrix planes. In this case the second phase has a spacing between platelets of less than 500 Å and the thickness of the platelets appears to be in the order of 100 Å.

It cannot be stated with assurance* that the structure *observed with electron microscopy* is that responsible for the extra reflections shown in the X-ray photograph or for the filamentary behavior in the superconducting measurements. However, the correlation is strikingly good and we are assured that the interstitial solid solution is certainly not a homogeneously random distribution under the techniques we have employed.

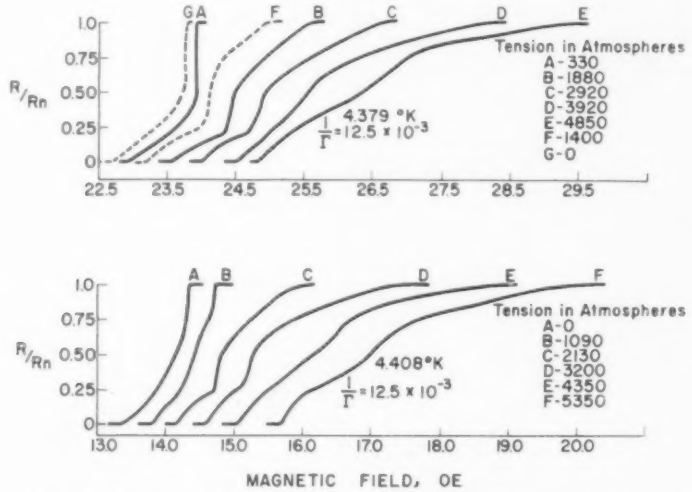


FIG. 8. Resistance transitions for tantalum under the indicated stresses in tension.

INHOMOGENEOUS ELASTIC STRAIN

Several experiments were performed to determine the effect of inhomogeneous elastic strain on the resistance transition. The experiments are simply elastic tension experiments on polygrained specimens. Since Young's modulus for tantalum is anisotropic[18], the strain in individual grains is

* The thermal treatment is of significance in determining the size of the platelets. The thin foils quench much faster by radiation than thin wires. Consequently a direct comparison of specimens with identical histories is difficult.

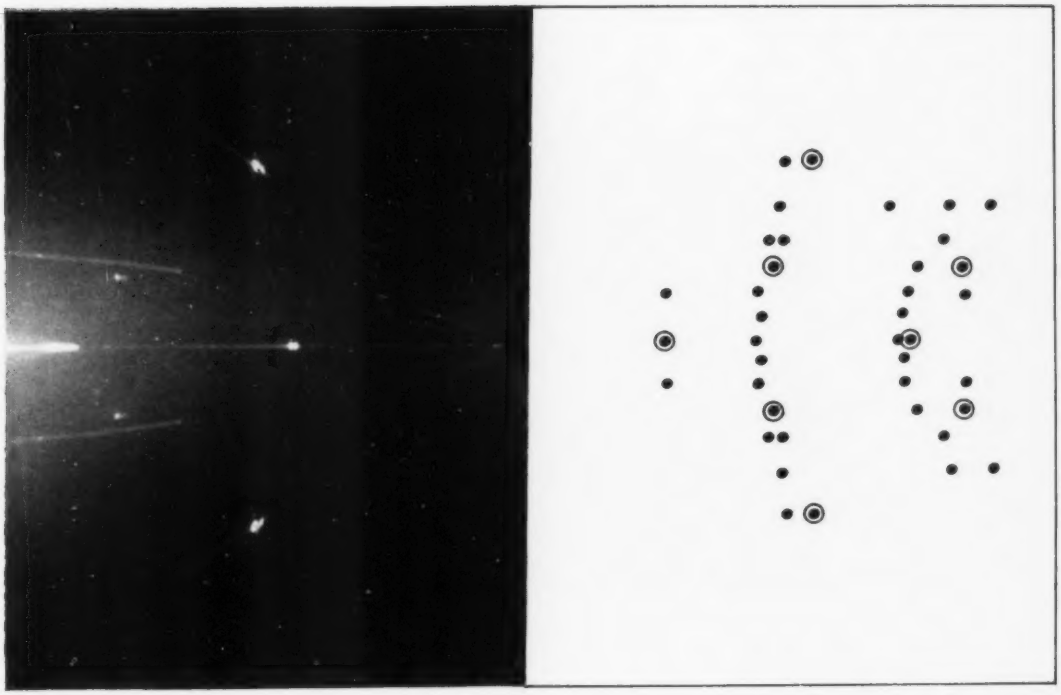


FIG. 6. X-ray photograph and tracing of the photograph showing super lattice lines for tantalum with absorbed nitrogen.

[facing p. 374

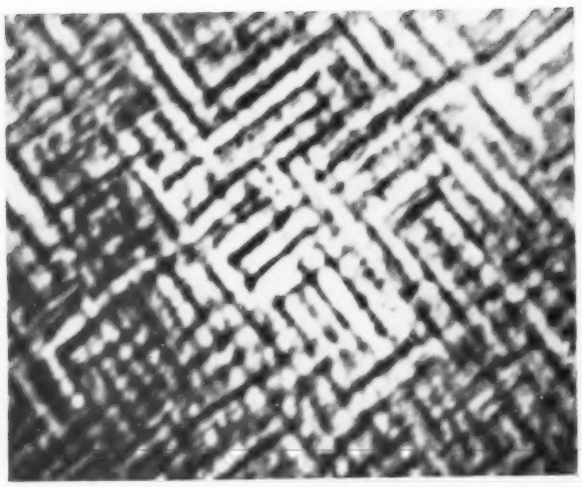


FIG. 7. Widmanstätten pattern for thinned tantalum foil with absorbed nitrogen (electron transmission × 65,000).

expected to be a function of the orientation of the grain and also a function of the constraints set up by neighboring grains. Homogeneous hydrostatic tension should have an effect opposite to the pressure effect as measured by JENNINGS and SWENSEN[19]. That is, the critical field should increase by $0{\cdot}285\times10^{-3}$ oersted-atm of tension* if shear, as previously shown[20], is not effective in changing H_c or T_c.

The magnitude of the effect shown in Fig. 8 is much greater than predicted. This may be explained by the fact that the resistance transition is characteristic of that part of the specimen with the highest critical field and the fact that a high critical field in certain regions can be produced by the inhomogeneous nature of the strain. The broadening of the transition also may be explained in this manner and an estimate can be obtained indicating that the largest hydrostatic tension stress is approximately eleven times the average resolved value ($1/3\ \sigma_T$).

The effect of tension, like the effect of pressure, appears to be reversible and independent of temperature, at least for the two temperatures used in Fig. 8. The reversibility is demonstrated by transitions F and G in Fig. 8(a) where these transitions were taken after transition E at 4850 atm. It may be noted that the specimen employed for these experiments was not pure ($1/\Gamma = 12{\cdot}5\times10^{-3}$). In fact the specimen was purposely chosen with sufficient interstitial nitrogen to pin down dislocation motion and hence to prevent the irreversible contribution from plastic flow. An effect of similar magnitude was obtained by employing a pure specimen ($1/\Gamma < 2\times10^{-3}$) after the metal was work hardened at 4°K.

PLASTIC DEFORMATION

The effect of cold work on the resistance transition depends largely upon the temperature at which damage is created and upon the recovery and recrystallization temperature for the particular metal under consideration. In the case of pure indium, deformation at room temperature has no significant effect[4] since recrystallization occurs at a lower temperature than ambient. Tantalum however has a recrystallization temperature in the region of 1000°C and can sustain a significant density of defects at room temperature.[21,22] The effect on the resistance transition at 4·25°K as a function of the increase in $1/\Gamma$ is shown in Fig. 9.

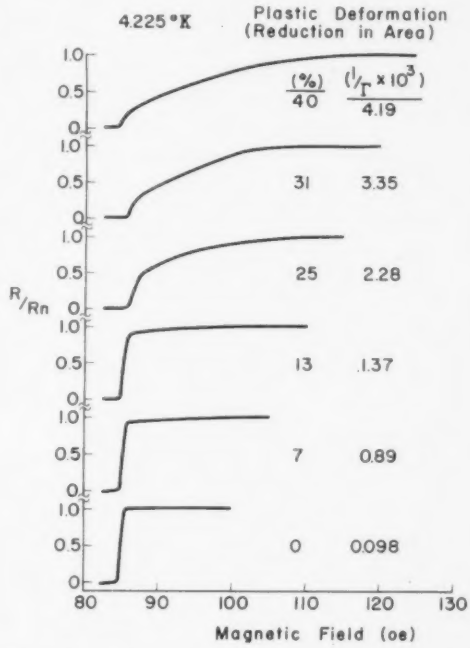


FIG. 9. Resistance transitions for cold-worked tantalum with the indicated reduction of area and resistance ratio.

It will be noted that the transition begins at the same magnetic field independent of cold work, while the end of the transition occurs at a magnetic field which increases with $1/\Gamma$. The broadening of the transition in magnetic field amounts to approximately 40 oersteds for the most heavily deformed specimen. The transition is thus eight times broader than the transitions shown in Fig. 8 for specimens with inhomogeneous stress of the order of 20,000 atm. Therefore, to explain the broadening due to cold work in terms of residual stresses, these stresses must be in the order of 10^5 atm. It is indeed unlikely that residual stresses of this magnitude are present. One significant difference in the broadening of transitions due to cold work, as opposed to inhomogeneous strain effects, which are essentially independent of

* Tensile stress σ_T can be resolved into a hydrostatic tension component of $1/3\ \sigma_T$ and into two pure shear components.

temperature, is the decrease in broadening, to only a few oersteds near T_c. We are therefore tempted to look for an explanation similar to that which will explain the effects of impurities.

In the range of 80°K and lower a significant density of point defects can be created by plastic deformation and held in the lattice. For many metals[22] T_c increases with plastic deformation at these low temperatures and the effect disappears with annealing below their recrystallization temperatures. Point defects appear to be responsible for changing the critical temperature since these defects anneal out of a metal prior to recrystallization.

DISCUSSION

It has been demonstrated that three kinds of lattice defects have significant effects on the characteristics of the resistance transition in tantalum. Interstitial impurities, particularly nitrogen, develop a filamentary substructure which appears to have distinct superconducting properties, e.g. a critical field H_n independent of composition up to approximately 0·20 atomic per cent. The amount of substructure, as evidenced by its ability to carry a supercurrent of high density, qualitatively increases with increasing nitrogen content. Even so, flux measurements indicate that penetration is essentially complete at magnetic fields lower than H_n. The metal therefore appears to be breaking up into a fine intermediate state. The X-ray measurements and the electron microscope show that the inhomogeneous phase mixture is not necessarily limited to low temperatures. Indeed the presence of a second phase in some specimens is confirmed by X-ray techniques and a second phase is also found in thinned foils of the same composition by electron transmission microscopy.

There are however many details which are not understood. One is that the second phase does not appear for some specimens submitted to X-ray analysis. Another is that ageing is not necessary to establish the filamentary behavior in resistance transitions while, in contrast, the Widmanstätten structure appears only with time in the electron beam. The spacing and thickness of the platelets may be too small to allow resolution in some of the X-ray work as is shown by the electron transmission study which does resolve the fine structure. These and other questions will be discussed in a subsequent publication on the metallurgy of dilute Ta–N solid solutions.

The filamentary behavior in impure tantalum is in some respects different from the effects found in impure lead or in other substitutional solid solutions. Specifically, a second phase is unlikely in many of the dilute substitutional alloys which do show filamentary behavior. Alternatively there is the possibility that short-range (i.e. of the order of the penetration depth) fluctuations in composition are present in these solutions which may arise from the formation of impurity atmospheres around dislocations (see SHAW[5]). A break-up into an intermediate state is then possible and broad irreversible flux transitions may be observed.

In any case it is important to realize that the resistance transition appears at the end of the flux transition and the two transitions may be widely separated. The broad continuous resistance transition is not related to the flux transition in any simple manner, and in most cases merely indicates a break-up into a fine intermediate state which generally occurs after a complete flux penetration at much lower field. Fortunately the similarity principle is obeyed and in those cases where only resistance transitions can be measured, a check on the validity of such measurements can be made by comparing them to the critical-field curves for pure bulk material.

The effect of inhomogeneous strain is to broaden the transition similarly at all temperatures. Plastic deformation broadens the transition in a qualitatively different manner; at low temperatures the effect is large in magnitude while near T_c the effect essentially disappears. This qualitative difference in the behavior leads the author to suggest that (in the case of plastic deformation) residual stresses alone are not responsible for the broadening of the transition. Indeed the situation seems to be quite complex and may require an analysis which involves other changes in the properties of the metal.

Inhomogeneous strain can be expected to be present in any polycrystalline specimen which displays an anisotropy in the linear expansion coefficient when the specimen is cooled to liquid-helium temperatures. This effect also can be expected to be present in any polycrystalline material with anisotropy in the elastic constants

if the material is under a tensile stress. The latter condition is adequately met in a thin film which exhibits a different coefficient of expansion from the substrate to which it adheres. Shear stresses will also be present in such a film and, provided the critical shear stress for plastic flow is sufficiently low, some irreversible effects can be present. However the relaxation in shear stress may result in a plastic strain less than the strain produced by thermal expansion. Plastic strains are generally small and are likely to have insignificant effects on the superconducting transitions. For extremely thin films the critical shear stress may increase by orders of magnitude and the internal shear stresses cannot then be relaxed.

Acknowledgements—The author is grateful to D. T. NOVICK and to D. J. QUINN who assisted with the experiments in superconductivity, to N. STEMPLE who did the X-ray work and to D. CAMERON who helped with the electron microscopy. Discussions with many persons in the IBM Cryogenics Department have been helpful. The author is especially indebted to R. A. CONNELL and W. B. ITTNER, III for their continued interest in this work and to Y. BUDO for his able assistance with the apparatus.

REFERENCES

1. D. P. SERAPHIM and R. A. CONNELL, *Phys. Rev.* **116,** 606 (1959).
2. K. MENDELSSOHN, *Proc. Roy. Soc.* A**152,** 34 (1935).
3. P. R. DOIDGE, *Phil. Trans.* A**248,** 553 (1956).
4. M. D. REEBER, *Phys. Rev.* **117,** 1476 (1960).
5. D. E. MAPOTHER, D. L. DECKER and R. W. SHAW. *Phys. Rev.* **112,** 1888 (1958); D. L. DECKER, Thesis, University of Illinois (1958); R. SHAW, Thesis, University of Illinois (1959).
6. W. B. ITTNER, III, and J. F. MARCHAND, *Phys. Rev.* **114,** 1268 (1959).
7. T. E. FABER, *Proc. Roy. Soc.* A**214,** 392 (1952); C. J. GORTER, *Physica, 's Grav.* **2,** 449 (1935); H. LONDON, *Proc. Roy. Soc.* A**152,** 560 (1935).
8. F. LONDON, *Superfluids* Vol. 1. Wiley, New York (1950).
9. D. WHITE, C. CHOU and H. L. JOHNSON, *Phys. Rev.* **109,** 797, 788 (1958); A. WEXLER and W. S. CORAK, *Ibid.* **85,** 85 (1952).
10. D. P. SERAPHIM, J. I. BUDNICK and W. B. ITTNER, III, *Trans. Amer. Inst. Min. (Metall.) Engrs.* **218,** 527 (1960).
11. J. I. BUDNICK, *Phys. Rev.* **119,** 1578 (1960).
12. D. P. SERAPHIM, D. T. NOVICK and J. I. BUDNICK. Submitted to *Acta Met.*
13. E. GEBHARDT, H. SEGHEZZI and W. DÜRRSCHNABEL, *Z. Metallk.* **49,** 577 (1958).
14. E. GEBHARDT and H. SEGHEZZI, *Z. Metallk.* **50,** 52 (1959).
15. J. BARDEEN, L. N. COOPER and J. R. SCHIEFFER, *Phys. Rev.* **108,** 1175 (1957).
16. E. A. LYNTON, B. SERIN and M. ZUCKER, *J. Phys. Chem. Solids* **3,** 165 (1957).
17. G. CHANIN, E. A. LYNTON and B. SERIN, *Phys. Rev.* **114,** 719 (1959).
18. D. I. BOLEF and M. MENES, *Bull. Amer. Phys. Soc.* **5,** 40 (1960).
19. L. D. JENNINGS and C. H. SWENSON, *Phys. Rev.* **112,** 31 (1958).
20. P. FISKE, *Phys. Rev.* **96,** 821 (1954). (Abstract)
21. M. P. LUCASSON. *Séance Du* **12,** 796 (1957).
22. V. I. CHOTKEVICH and J. R. GOLIK, *J. Exp. Theor. Phys.* **20,** 427 (1950).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 378–380. Printed in Great Britain

USE OF A SUPERCONDUCTING TRANSMISSION LINE FOR MEASURING PENETRATION DEPTHS

D. R. YOUNG, J. C. SWIHART, S. TANSAL and N. H. MEYERS

Research Laboratory, International Business Machines Corporation, Poughkeepsie, New York

I. INTRODUCTION

IF A transmission line is constructed using superconducting films for conductors and an insulation layer comparable in thickness to the superconducting penetration depth, then the superconducting penetration effect markedly influences the propagation constants of the line. The use of a

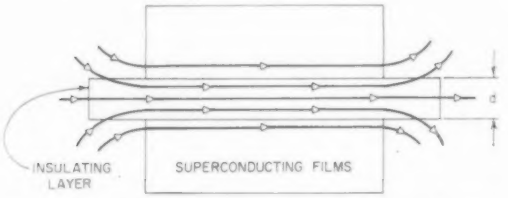


FIG. 1. Magnetic-field penetration in superconductors.

transmission line for measuring penetration depth was first suggested by PIPPARD[1]. The situation is illustrated by Fig. 1.

Since the electrostatic charges are located on the inside surfaces of the films the capacity per meter is given by

$$C = \epsilon \frac{W}{d} \text{ (m.k.s. units)}$$

If the penetration were zero the inductance would be

$$L = \mu_0 \frac{d}{W}$$

However, if the superconducting film thickness is large compared to the penetration depth London's equation yields

$$L = \mu_0 \frac{d}{W}\left(1 + \frac{\lambda_1}{d} + \frac{\lambda_2}{d}\right)$$

where λ_1 and λ_2 are the penetration depths for films no. 1 and 2 respectively. The correction terms λ_1/d and λ_2/d are each composed of two equal terms $(\lambda/2d)$ due to the magnetic-field energy and the kinetic energy of the electrons. If the superconducting films are reduced in thickness to become comparable to the penetration depth then the kinetic-energy term becomes large compared with the magnetic-field energy term since the kinetic energy depends upon the current density squared. The correction terms become

$$\frac{\lambda_1}{d}\coth\frac{l_1}{\lambda_1} \text{ and } \frac{\lambda_2}{d}\coth\frac{l_2}{\lambda_2}$$

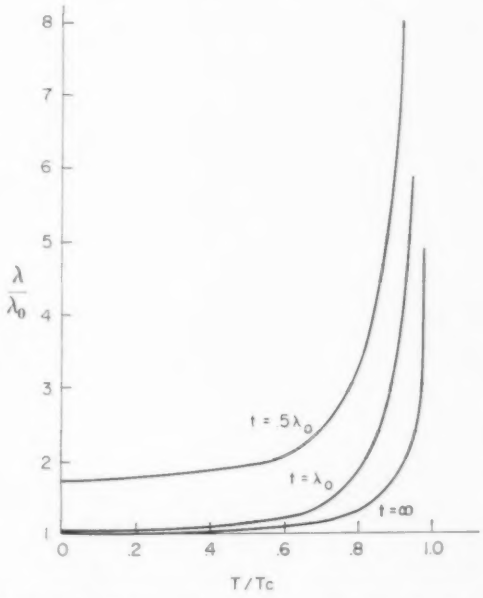


FIG. 2. Variation of effective penetration depth with temperature for several film thicknesses.

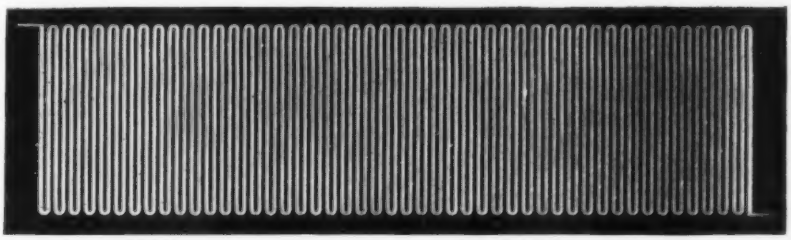


FIG. 4. Glass mask used for evaporating transmission line.

FIG. 6. Photograph of oscilloscope trace showing signal reflections. ($T = 3{\cdot}6485^{\circ}$K; pulse duration = 9 mμsec; pulse amplitude = 300 mA; deflection sensitivity = 7·1 V/div.; sweep speed = 20 mμsec/div.)

[*facing p.* 378

where l is thickness of the corresponding film. This effect increases the "effective λ" as measured for thin films, as shown by Fig. 2. The expected increase in L for thin films indicates a new type of limitation in circuit speed to be encountered for thin films.

The velocity of propagation is given by

$$v = \frac{1}{\sqrt{(LC)}}$$

and the characteristic impedance by

$$Z = \sqrt{(L/C)}$$

It can be seen that the penetration effect decreases the velocity of propagation and increases the characteristic impedance. Measurements of these quantities will enable us to calculate the penetration depth. If C is known then velocity measurements are sufficient.

II. EXPERIMENTAL TECHNIQUES

The actual line being used is 50 in. long and consists of a zig-zag superconducting tin film on top of, but insulated from a superconducting lead ground plane. This geometry is used to eliminate registration requirements that would be encountered if two zig-zag lines were used. The line is evaporated through a glass mask obtained from The Corning Glass Company and made by their photo-etch process. Coupling effects between adjacent segments are negligible due to the extremely thin insulating layers used. The construction of the line is shown in Fig. 3 with the associated magnetic-field configuration. The glass mask used is shown in Fig. 4 and the circuit in Fig. 5.

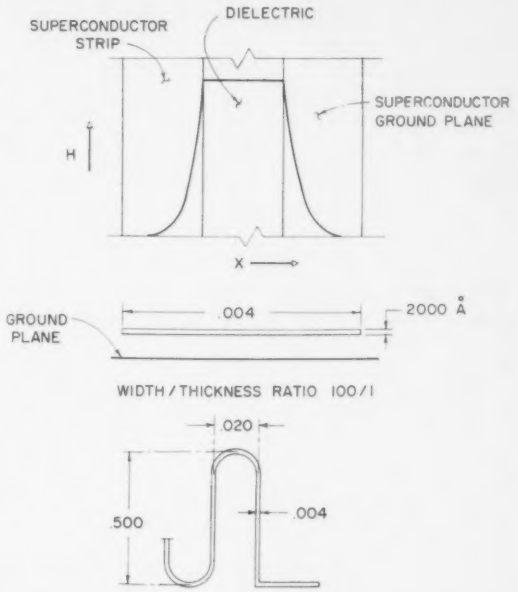


FIG. 3. Construction of delay line with magnetic-field configuration.

The input and output impedances are 50 Ω. Since the line has a characteristic impedance of about 1 Ω, the pulse is able to traverse the line many times before it is attenuated. In the normal state the signal is attenuated very rapidly even

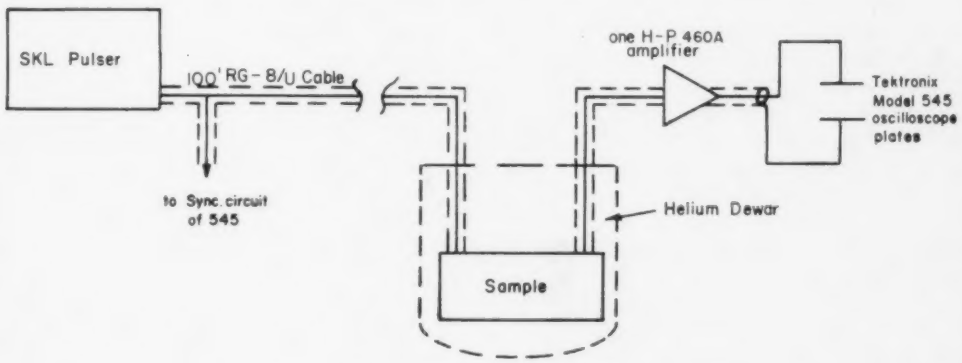


FIG. 5. Block diagram of circuit.

before it traverses the line once. It has been found that the loss in the signal (in the superconducting state) is only that due to the terminations and that there is no appreciable loss in the line itself. Fig. 6 is a photograph showing the reflections of the signal as observed on the oscilloscope. The velocity is measured using the known sweep speed of the oscilloscope.

III. RESULTS

Only preliminary results have been obtained to date. These are shown in Fig. 7. The dependance of the velocity on temperature is clearly observed. The data suggests a penetration depth much larger than the bulk values reported earlier. It will be necessary to improve the accuracy of the thickness measurements and dielectric-constant measurements before these results can be taken seriously. However, it should be pointed out that these results are not inconsistent with ITTNER's[2] calculations showing that λ is expected to be larger for thin films. The pulse shape is maintained even after many reflections so that it can be said that there is no dispersion observable in the frequency range of this experiment (up to 100 Mc/s).

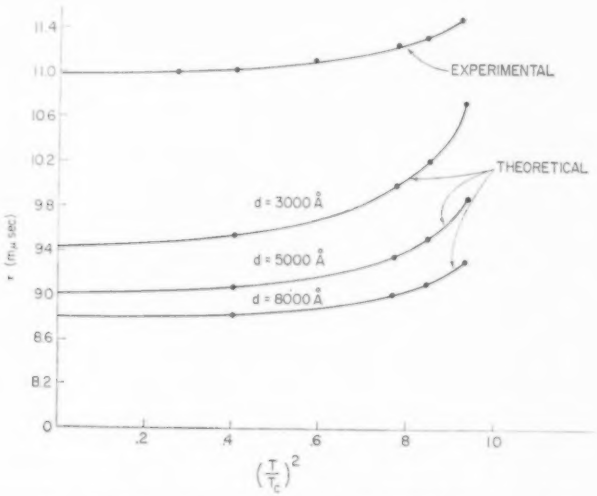


FIG. 7. Delay time as a function of $(T/T_c)^2$ for several film thicknesses.

IV. DIRECTION FOR FUTURE WORK

Since the change in velocity due to changes in penetration depth is reasonably large as regards percentage due to the thin insulating layers used, this technique should enable us to study small effects such as the dependence of λ on the magnetic field. To make this possible the experiment is being modified to use the line short circuited at one end and open at the other as a quarter-wave resonant circuit. The velocity will then be determined by a measurement of the resonant frequency. This should increase the accuracy of the measurement.

REFERENCES

1. A. B. PIPPARD, *Proc. Roy. Soc.* A**191,** 399 (1947).
2. W. B. ITTNER, III, *Phys. Rev.* **119,** 1591 (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 381–387. Printed in Great Britain

EDGE EFFECTS IN SUPERCONDUCTING FILMS

RALPH B. DeLANO, JR.

International Business Machines Corporation, Research Laboratory, Poughkeepsie, New York

INTRODUCTION

THE thin-film cryotron was suggested several years ago in order to overcome the speed limitations of the wire-wound cryotrons by increasing the resistance of the gate. Most of the work has been with evaporated films where the technology was more of an art than a science. One of the major problems in this work has been an understanding of the factors which affect the reproducibility of metal films produced by evaporation. The major characteristics of interest here are the critical field as a function of temperature, the sharpness of the magnetic transition from the superconducting to the normal state, and the critical temperature. Since the edges of evaporated film can have a profound effect on the first two characteristics a comprehension of the edge effect is important in order to make reproducible devices.

Previous workers[1], believing that the width of the magnetic transitions of strip-shaped films was influenced by the penumbra at the edge, made their measurements on a disk-shaped sample with the electrodes a substantial distance from the edge so that only a very small portion, if any, of the measuring current flowed in the edges.

This paper discusses the effect that edges of typical evaporated tin films have on their magnetic and temperature transitions, and the importance of these effects on thin-film cryotron characteristics. Practical methods for eliminating the edge effect are mentioned. A description of the apparatus and techniques used will be found in previous papers.[2–4]

MAGNETIC TRANSITIONS

There are two reasons why the edges of a film can cause a broad transition. The critical field of a film is larger for thinner films because of penetration effects.[5] The sloping edges existing in the penumbra are thinner than the bulk of the film and the measuring current is forced into the edges. No resistance is measured until the field is sufficient to drive the edges normal and the width

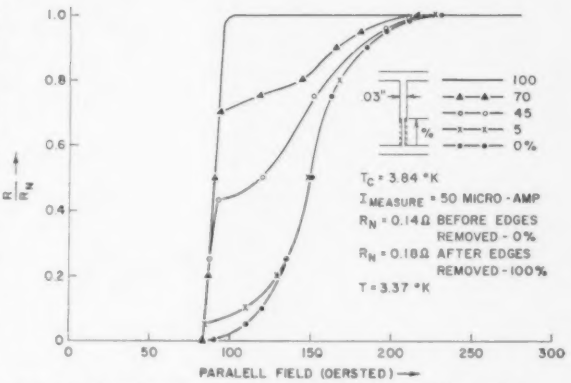


FIG. 1. Changes in the magnetic-field transition of a tin film as edges were removed.

of the transition is determined by the characteristics of the edges. In addition, imperfections tend to concentrate in the film edges, as discussed in the next section, and can cause a broadening of the transition.

The effect of successively removing increasing portions of the edges on the magnetic transition of a tin film is shown in Fig. 1. In this case the edges were removed by mechanical trimming. Etching with an acid in glycerin has also been used to remove the edges.

Etching has further shown that a uniform grain size is desirable if narrow transitions are required. Films having large grain size and those having small grain size can both have sharp transitions. When large and small grains both occur in the same film, however, the transitions can be nearly an order of magnitude wider than when the grains are uniform.[2]

Thin films will melt at a lower temperature than thick films,[6] and this effect can be used to cause the edges to break up and become discontinuous.[7] There is in this technique some danger that, if the edges contain more imperfections than the bulk of the film, imperfections may diffuse into the bulk of the film and cause deterioration of the transitions.

Since the transitions measured are those of the thinner edges, the critical field is reduced considerably by removing them. The change in critical field as a function of temperature is shown in Fig. 2 for a tin film 3800 Å thick. The solid curves are

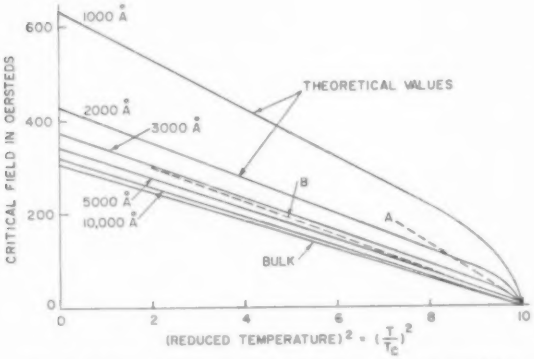


FIG. 2. Change in the critical field of a tin film resulting from removing the edges. (a) Before removing edges. (b) After removing edges (thickness 3800 Å).

obtained by applying London-type corrections to the bulk values for tin. The critical field for some films having a large penumbra has been reduced by an order of magnitude by removing the edges.

Films can be deposited in a manner that makes them free of the edge effect. This is accomplished by increasing the mobility of the atoms on the substrate so that what would normally be the thin film edge is now a series of islands without continuity from one particle to the other. Heating the substrate during deposition will cause the edges to break up in this manner.[3] Depositing the films in a system which is baked and free of water vapor, oxygen and hydrocarbons will also casue the edges to break up.[8]

TEMPERATURE TRANSITIONS

Since variations in the critical temperature of a film depend on imperfections and not on film thickness, a study of the temperature transitions along with the magnetic transitions makes it possible to determine whether the edge effect is caused by imperfections, the thinness of the edge, or both.

The temperature transitions of a film are shown in Fig. 3 before and after the edges were removed. This film had a wide penumbra because it was deposited at a high deposition rate with the substrate close to the source. Trimming produced a large reduction in the critical temperature of the transition taken with a small measuring current. This demonstrates that there was a large concentration of imperfections concentrated in the edges. The electron mean free path computed at 4·2°K for this film was 0·09 cm, which is two orders of magnitude larger than that for films deposited at conventional rates and normal substrate-to-source distances. Although the edges of the film had a high concentration of imperfections, the bulk of the film was exceptionally good.

Here, as in magnetic transitions of tin films, the small measuring current flows in the edges, since the edges have a higher critical temperature than the bulk of the film. Consequently the sharpness of the transition is quite sensitive to the characteristics of the edges.

The concentration of impurities is not as high in the edges of films deposited at rates less than 100 Å/sec and with source-to-substrate distances of several inches. The temperature transition of such a film is shown in Fig. 4 before and after removing the edges. The bulk portion of this film had a rather large concentration of imperfections since the electron mean free path computed at 4·2°K was $3{\cdot}7 \times 10^{-4}$ cm. The change in temperature transition is not as dramatic as the previous film indicating that the imperfections are more uniformly distributed. The temperature transition is about three times sharper than that for the previous film showing that uniformity rather than amount of imperfections determines transition width.

Before the edge effect was understood attempts were made to improve the sharpness of the transitions by annealing. While annealing did reduce the residual resistivity, indicating a reduction in imperfections, the transitions did not improve. Sometimes they became worse.

By measuring the temperature transition of a film, before and after annealing, with both small and large currents, the effect of film edges was observed. The results of such an experiment are plotted in Fig. 5. The ratio of the room-temperature to the liquid-helium-temperature resistance for this film improved from 38 to 163 but the sharpness of the transition did not improve. After annealing, the transition measured with a small current occurred at a lower temperature while that measured with a large current occurred at a higher temperature. Since the small-current transition is determined primarily by the edges, the lowering of the transition temperature indicated that the annealing removed some of the imperfections from the edges.

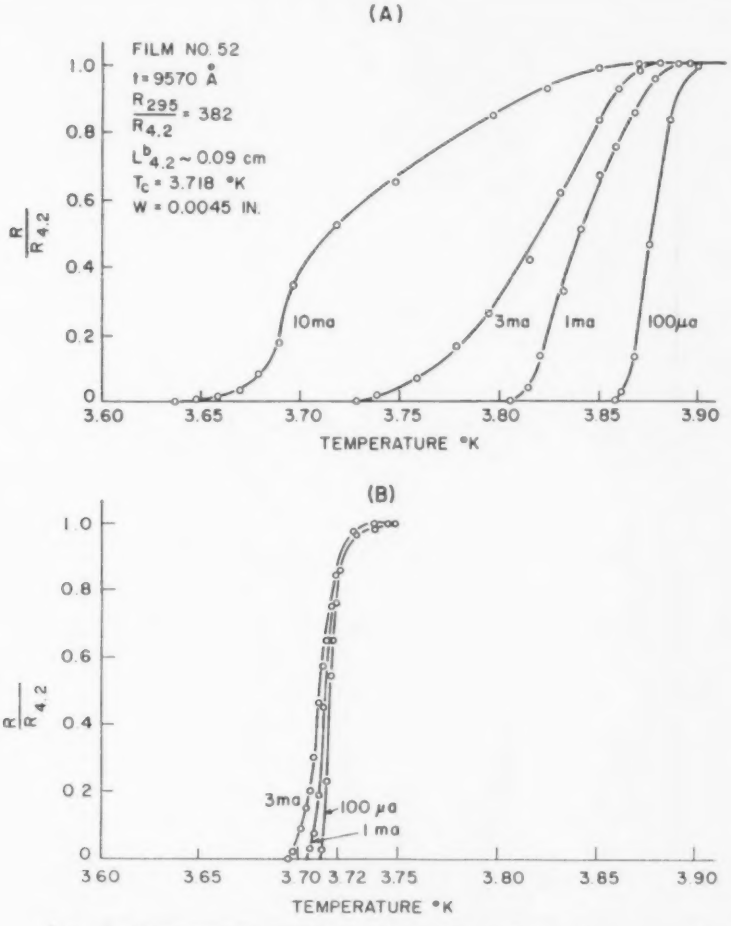


FIG. 3. Temperature transitions for a film having a large penumbra showing the effects of removing the edges. (a) Transitions for various measuring currents before the edges were removed. (b) Transitions with both edges completely removed.

The increase of the critical temperature measured with large currents observed after annealing is also significant. The superconducting current-carrying capacity of the edges had been decreased during the annealing as shown by the lowering of the critical temperature when measured with

a small current. Consequently the effect of the edges on the large-current transition would have been to lower this transition. Since it occurred at a higher temperature the possibility arises that the imperfections which were driven out of the edges diffused into the bulk of the film.

provement produced by the anneal was small. The transition temperature measured with a large current did not increase as it did when the previous film was annealed before removing the edges. The transitions of the film annealed after removing the edges are sharper. Consequently

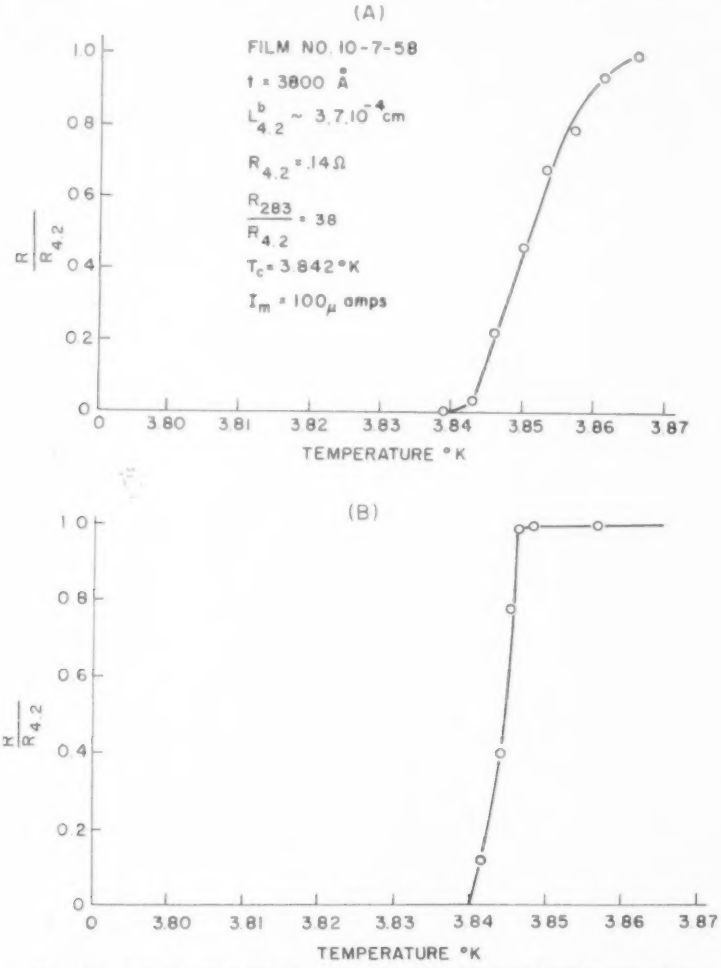


FIG. 4. Effect of removing the edges of a film deposited under normal conditions. (a) Temperature transition before removing the edges. (b) Edges removed.

This possibility was tested further by annealing a film after the edges had been removed by etching. Fig. 6 shows the temperature transitions taken at small and large measuring currents. The resistance ratio of this film was high and the im-

the conclusion that imperfections can diffuse from the edge into the bulk of the film seems probable.

CRYOTRON CHARACTERISTICS

The reduction in critical field resulting from

removal of the edges will increase the current amplification of a cryotron. The current amplification is the ratio of the critical current of the gate to the current in the control required to produce the critical field in the gate. If one cryotron determined by the mask techniques used to make the control, which is the narrowest line and the ratio of gate width to control width necessary to achieve the desired amplification. Therefore the ability to decrease the width ratio and still obtain

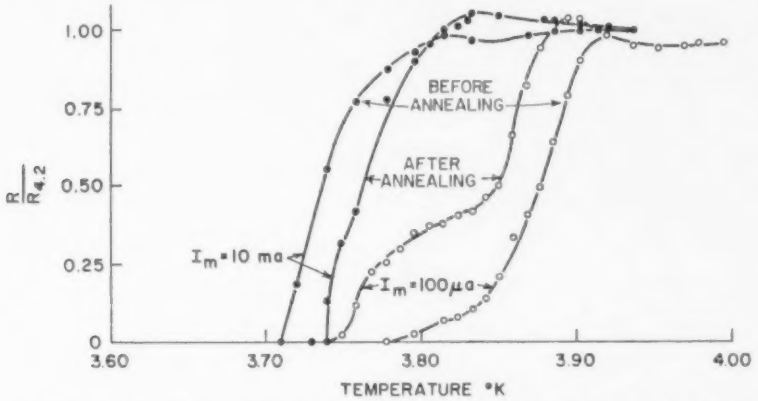


FIG. 5. Effect of annealing on the temperature transition of a film with edges.

flip-flop is to drive another the operating amplification must be greater than 1. The critical current is probably porportional to the critical field, but since the greatest portion of a large current flows in the bulk of the film the value of critical field of interest is that for the bulk of tin film, rather than the edge. Consequently trimming the edge of a film does not affect the critical current appreciably but may greatly reduce the critical field and consequently increase the amplification.

A number of experiments have been performed to verify that removing the edges does not appreciably reduce the critical current. For instance removing the edges of a film having a wide penumbra reduced the critical field by a factor of 20, yet the critical current was reduced by only 25 per cent.

One consequence of being able to obtain increased amplification for a given width ratio is that smaller cryotrons can be made, since a certain amplification is usually desired. The field produced by the control of a cryotron is inversely proportional to the width of the control, and the critical current is proportional to the width of the gate; consequently the amplification is proportional to the ratio of the width of the gate to that of the control. Minimum cryotron size is the necessary amplification, results in a smaller cryotron.

Sharper magnetic transitions will increase the operating speed of cryotron circuitry. The pulses which appear across the input terminals of a cryotron will have a finite rise time if they come from the output of a previous cryotron. The amount of resistance introduced in the driven circuit will be proportional to the rate of rise of the driving pulse and the sharpness of the magnetic transition. Since the speed is proportional to the amount of resistance introduced, sharp transitions will result in faster circuitry.

If the cryotrons are biased from an external source, sharper transitions will make a greater reduction of the working current possible. Consequently, in circuits where heating limits the operating speed, improving the transition sharpness will result in higher speed and closer packing of cryotrons.

Elimination of the dependence of transitions on the characteristics of film edges will also improve reproducibility. The edges are more sensitive than the rest of the film to such variables as substrate cleanliness, substrate temperature, mask-to-substrate spacing and residual gas pressure. Each of these variables is difficult to control.

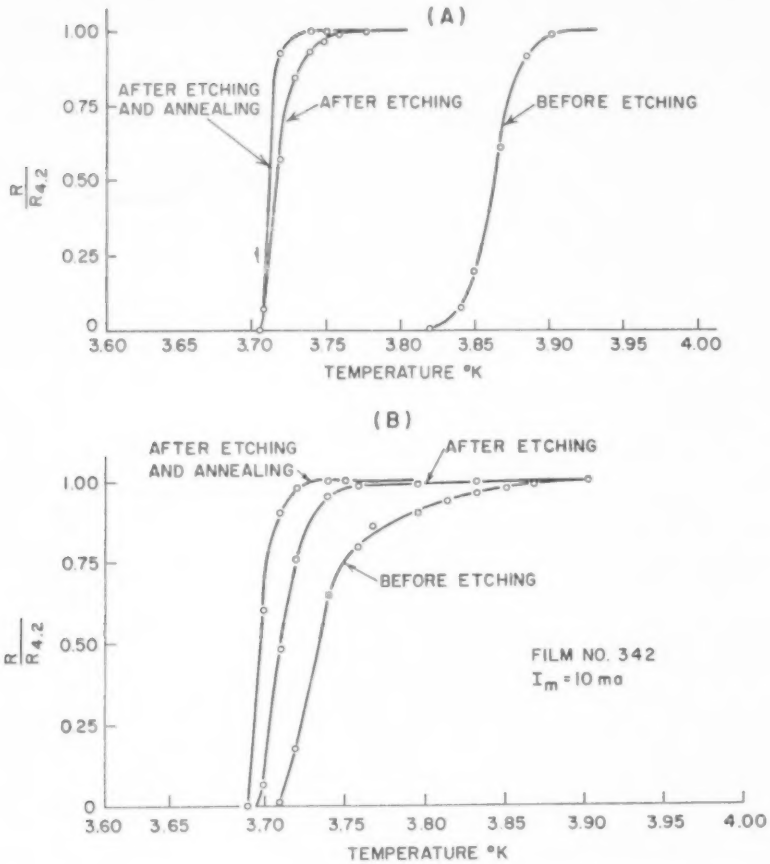


FIG. 6. Effect of annealing after removing the edges: (a) Measuring current of 100 μA. (b) Measuring current of 10 mA.

CONCLUSIONS

This work shows that edges play an important role in the technology of the evaporated-film cryotron. The broadening of the magnetic transition is caused by the sloping edges, and by the imperfections that can concentrate in the edges during the evaporation process. Several methods of removing edges or otherwise eliminating the edge effect have been mentioned.

The elimination of the edge effect improves cryotron performance by increasing the operating speed and reducing the size. An appreciable advance in the reproducibility of cryotron characteristics should also be obtained.

Acknowledgements—The author has reported the results of experiments performed by a number of people. Many of the experiments were performed by G. J. KAHAN. Others were done by A. E. BRENNEMANN, H. L. CASWELL and R. T. C. TSUI, all of the Poughkeepsie Research Laboratory, and M. E. BEHRNDT, R. H. BLUMBERG and G. R. GIEDD of the Federal Systems Division Development Laboratory.

REFERENCES

1. E. T. S. APPLEYARD, J. R. BRISTOW, H. LONDON and A. D. MISENER, *Proc. Roy. Soc.* A**172,** 540 (1939).
2. G. J. KAHAN, R. B. DeLANO, JR., A. E. BRENNEMANN and R. T. C. TSUI, *IBM J. Res. Dev.* **4,** 173 (1960).

3. M. E. BEHRNDT, R. H. BLUMBERG and G. R. GIEDD, *IBM J. Res. Dev.* **4**, 184 (1960).
4. H. L. CASWELL, *IBM J. Res. Dev.* **4**, 130 (1960).
5. W. B. ITTNER, III, *Proceedings of the Symposium on Superconductive Techniques for Computing Systems, Washington, D.C.*, 1960, p. 239. Office of Naval Research (1960).
6. M. BLACKMAN and A. E. CURZON, *Structure and Properties of Thin Films* p. 217. Wiley, New York (1959).
7. H. L. CASWELL and C. CHIOU. Unpublished work.
8. H. L. CASWELL, *Proceedings of the Symposium on Superconductive Techniques for Computing Systems, Washington, D.C.*, 1960, p. 262. Office of Naval Research (1960).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 388–398. Printed in Great Britain

OPERATION AND ANALYSIS OF PLANAR CRYOTRONS AND SIMPLE CRYOTRON CIRCUITS

G. B. ROSENBERGER

International Business Machines Corporation, Federal Systems Division, Kingston, New York

Abstract—Simple cryogenic loops containing two drive cryotrons and a number of sensing cryotrons have been constructed. It is possible with such loops to study both the operating characteristics of the individual cryotrons as well as the static and dynamic switching characteristics of the loop. Some of the characteristics of the individual cryotrons and the reproducibility of cryotrons with these characteristics are discussed. The speed with which current can be transferred from one branch of the loop to another has been both measured experimentally and calculated theoretically from the dimensions and various measured parameters of the loop and its cryotrons. There is found to be good agreement between theory and experiment. The loops have been run dynamically in an effort to detect thermal effects associated with heat dissipation in the resistive cryotron elements. The upper limit to the dynamic operation is currently set by the viewing equipment and, to date, no adverse heating effects have been observed.

INTRODUCTION

MUCH work has been done in an attempt to show the feasibility of using bulk superconductors to fabricate current-switching devices. Circuits made with devices of this type, such as the wire-wound cryotrons, have severe operational speed limitations because of large L/R switching time constants. These speed limitations become less significant when the devices are made of evaporated films rather than bulk materials.

The planar cryotron used in these studies is shown diagrammatically in Fig. 1. The soft superconductor (gate) is placed between the control line and the ground plane and insulated from them by thin layers of evaporated silicon monoxide (SiO). The control line crosses the gate at an angle of 90°. This study includes cryotrons with gates between 0·4 and 1·0 μ thick and between 50 and 1500 μ wide, and with control lines between 0·5 and 1·0 μ thick and between 12·5 and 250 μ wide. The slide used in this study consists of two gates with two control lines crossing each of the gates, so that characteristics of the four cryotrons can be measured individually. Both the gates and controls are constructed in the form of four-terminal networks so that independent current and voltage terminals are available. The gates are connected to the lands with lead (Pb) films so that the gates can be made without any bends in the tin film, and also so that the entire length of tin is over the ground plane.

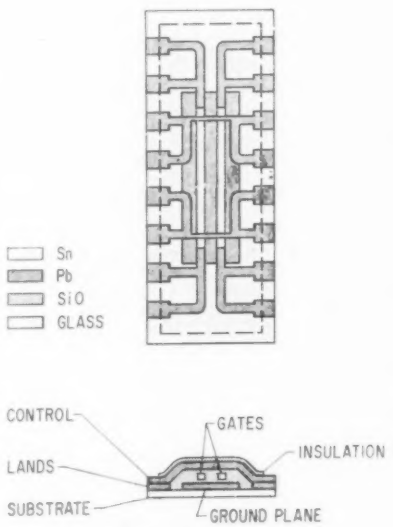


FIG. 1. Cryotron test slide.

CRYOTRON MAGNETIC-FIELD SWITCHING

When a current of the proper magnitude is

passed through the control film, the magnetic field of this control current destroys the superconducting state of the gate film. In order to determine the complete cryotron behavior it is necessary to know the magnitude of the magnetic field required to destroy the gate's superconductivity. This value of magnetic field, H_c^f, can be determined by means of an external magnetic-field measurement. The measured H_c^f is then compared to the calculated field impressed on the gate by current flow in the control line. H_c^f used in this study is the value of magnetic field required to restore one-half the normal resistance of the gate with effectively zero current through the gates. The transition width ΔH is defined as the change in applied field required to increase the measured resistance from 10 to 90 per cent of its full value. Pure bulk metals exhibit very sharp transitions, but in studies of the superconducting properties of evaporated films ΔH is frequently found to be quite large, i.e. up to 50 per cent or more of the value of H_c^f. Films with ΔH's that are sharp are desirable for the first studies of cryotrons and cryotron circuits because sharp transitions simplify both the tests and analysis of the test results. In addition, films with wide transitions are usually found to have values of H_c^f which are quite sensitive to the magnitude of the measuring current, while films with sharp transitions have critical fields which are relatively independent of the measuring current (in the range of low currents).

The field transitions, expressed in per cent of H_c^f, listed in Fig. 3 are typical of the results obtained on the cryotron gates used in this study. No difference could be detected in these transition widths for sense currents of 1 and 10 mA. It was found that sharp transitions could be produced by a number of techniques, one of which is discussed in a paper by BEHRNDT *et al.*[1]

It was found that H_c^f for two gates evaporated at the same time were usually identical. Plots of H_c^f as a function of temperature squared are shown in Fig. 2 for several samples made in separate evaporations. There is very little variation ($\pm 3{\cdot}3$ per cent) of the H_c^f values at low temperatures, but close to T_c the variation is generally much greater ($\pm$ 18 per cent). The critical temperatures for these films (listed in Fig. 2) indicate the typical variation in the critical temperature from sample

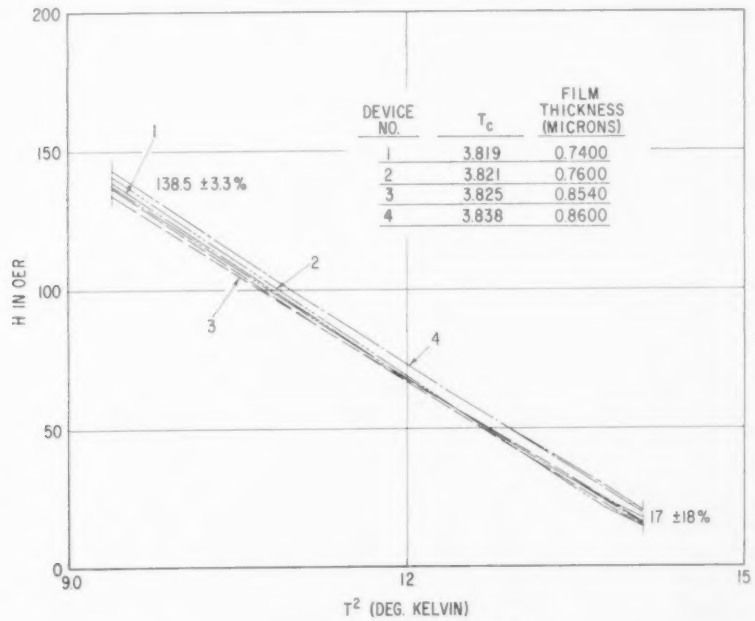


FIG. 2. Cryotron field results.

to sample. (T_c is defined here as the temperatures at which the gate resistance is equal to $1/2R_{4.2}$.) Fig. 3 illustrates these same magnetic-field transitions normalized for variations in critical temperature by plotting H_c' in terms of the reduced temperature $(T/T_c)^2$. The reproducibility is increased slightly near the critical temperature indicating that better control of T_c should yield even better reproducibility of H_c'. In summary, the magnetic-field transitions of the cryotron gates were found to be sharp, independent of small sense currents and reproducible.

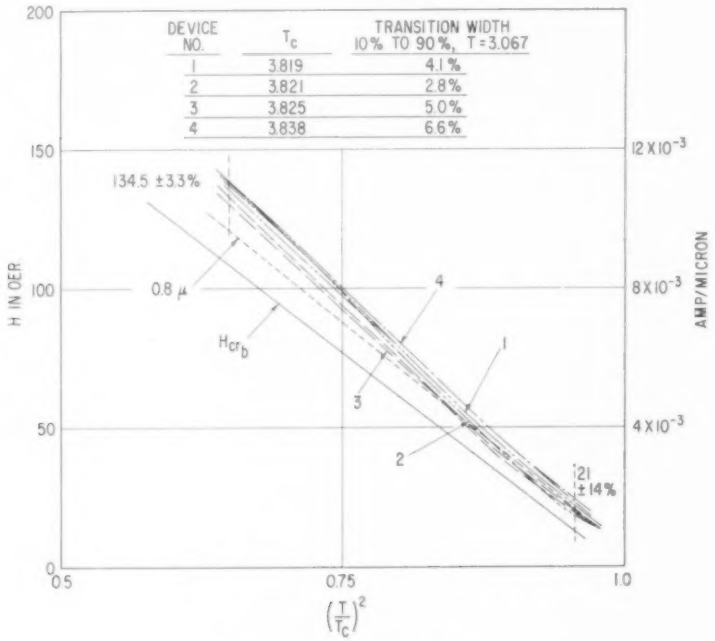


FIG. 3. Cryotron field results.

It is now necessary to determine the control current required to generate a field equal to the H_c' measured for a given sample. This control current, I_{co}, is determined by applying a small sense current to the gate and monitoring the voltage across the gate as the control current is slowly increased. A certain value of control current causes resistance to be switched into the gate (quite steeply for films with sharp field transitions) until the normal resistance of the portion of the gate under the control has been restored. Further increasing the control current results in very little additional resistance switched into the gate. I_{co} is arbitrarily defined here as the value of control current required to restore one half the full resistance. This definition is used to maintain consistency with the H_c' value used. Sharp transitions are advantageous for I_{co} measurements for the same reasons that they are advantageous for H_c' measurements.

Fig. 4 is a plot of I_{co} vs. T^2 for fourteen cryotrons evaporated on four different slides. On each of the four samples there was one cryotron which required higher control currents than the others. Optical microscope measurements proved these cryotrons to have wider control lines, and when the control-current values were normalized by plotting control current per unit control-line width, the magnitudes were brought back into line with the others. The range of values for these normalized control currents is plotted in Fig. 5 (shaded area).

The relationship between geometry and control current and its resulting magnetic field may be simply solved by approximating the control film

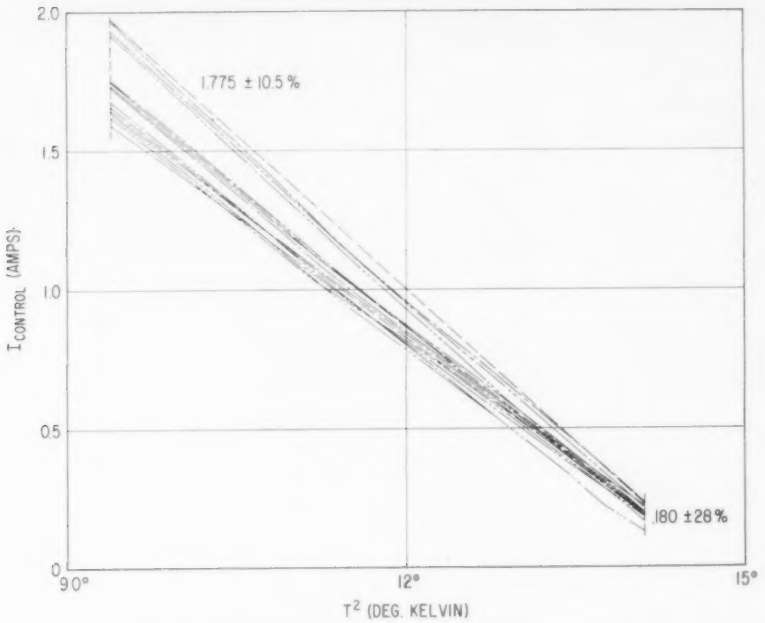


FIG. 4. Cryotron results I_{co} vs. T^2 for fourteen cryotrons.

and ground plane by two normal conductors with currents flowing in opposite directions.*

If the width of the control film is large compared to its spacing from the ground plane, Ampere's Law indicates that the horizontal component of magnetic field between the two conductors will be given from the expression, $I = Hw_c$, where I is the current in the control film, H is the magnetic field set up by this current in the small spacing between the ground plane and the control line, and w_c is the width of the control film. The value of field between the control line and the ground plane is nearly constant over the width of the control, and the field at all other points is effectively zero.

To verify experimentally the relationship, $I = Hw_c$, the critical magnetic-field values as measured by an external coil were averaged for all gates tested. This field value, plotted in Fig. 5, is expressed in amperes per micron of control width, so it represents directly the calculated I_{co} for a 1-μ line. The range of values obtained for cryotron control currents is also plotted in terms of unit control-line width (the shaded area shown in Fig. 5). These results show that the value of I_{co} calculated from the measured H_c' falls within the scatter of the experimental data, thus verifying the relationship.

CRITICAL CURRENT OF THE GATES (I_{gc})

Another parameter that must be ascertained in order to determine cryotron performance is the critical current of the cryotron gates. A four-terminal network is used to measure I_{gc} (see Fig. 1). The current is applied to the two most distant gate terminals and the critical current is recorded as the lowest current value that causes a voltage to appear across the two inner terminals.

The measured range of I_{gc} values for ten gates tested is plotted as a function of $(T/T_c)^2$ in Fig. 6. There is less than $\pm$ 10 per cent variation in the I_{gc} per unit gate width for films of at least 0·70μ thickness except for temperatures very close to T_c. If the magnetic field penetration into the gate film

* The model used can be pictured by taking the control film as one conductor and the image of the control film as the other. The net effect of the ground plane is approximated by assuming that any current applied to the control results in an equal and opposite supercurrent in the image control.

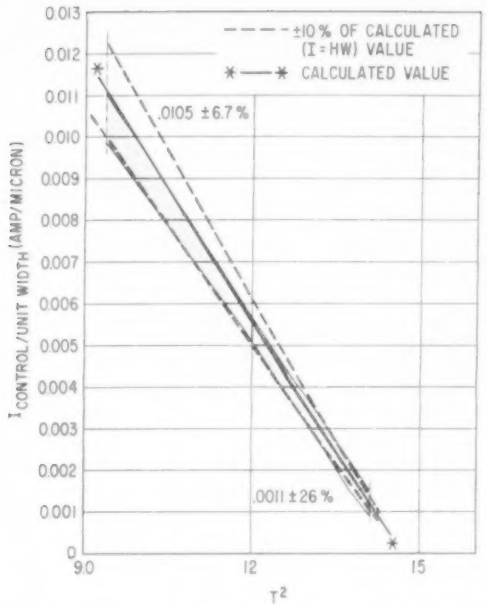


FIG. 5. Cryotron results. Control current per unit gate width vs. T^2.

is assumed to be negligible, the relationship $I_{gc} = H_c W_g$ becomes more nearly valid, and it thus may be expected to specify the upper limit of gate current to be found with extremely thick gate films. For thinner films where the penetration depth becomes comparable to the thickness of the film, the gate current would be expected to be somewhat less than the value for the thick films. The plots shown in Fig. 6 for thinner films fall below the maximum value, thus indicating that there is indeed a greater penetration for these films. Moreover, there is also a large degree of scatter for the thinner films so that serious difficulty may be encountered with the reproducibility of thinner gates. In summary: $I_{gc} = H_c W_g$ does appear to be the upper limit of critical-current relationship for films that are at least $0{\cdot}70\mu$ thick. Good reproducibility can be obtained for gates of this thickness.

CRYOTRON GAIN

Gain (G) has been defined as I_{gc}/I_{co}. For films that are thick enough so that they exhibit bulk magnetic-field characteristics, the expressions for I_{gc} and I_{co} are:

$$I_{gc} = H_c W_g \text{ and } I_{co} = H_c W_c$$

so that for films of this type $G = W_g/W_c$. Since the measured field H_c' for the gates tested is higher than H_c and since I_{co} is proportional to H_c', a more realistic expression for gain is $(H_c/H_c')(W_g/W_c)$ which is somewhat less than the ratio of the widths of the two films. Experimental values of gain for the thicker films were found to be between 0·75 and 0·90 W_g/W_c.

A minimum gain of unity is required to enable one cryotron circuit to drive a similar circuit.

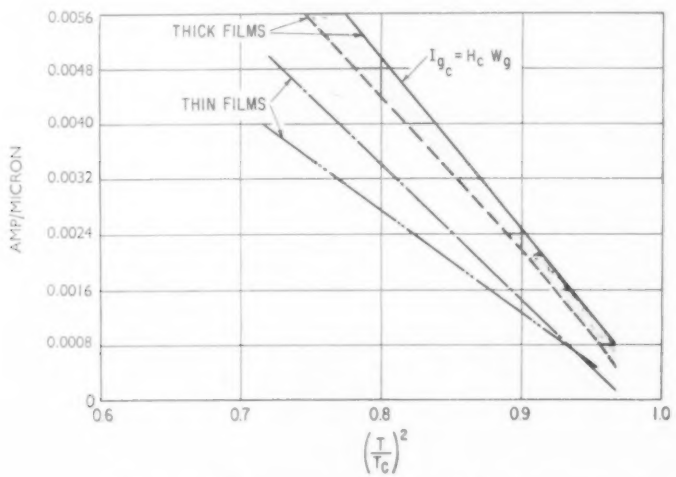


FIG. 6. Critical current per unit width for cryotron gates.

This means the I_{gc} must be at least as great as I_{co}. Since I_{gc} is proportional to gate width, high gains can be obtained by making the gates wide. However, the resistance switched into the gate film is inversely proportional to the gate width so that the increase in gain is obtained by making the L/R switching time greater. The circuit time constant will be given further consideration under the discussion of the flip-flop studies, but it should be noted here that the circuit designer will have to determine the proper compromise between gain and speed for each particular application.

According to London's theory, the gain should rise sharply with decreasing temperature until it reaches a maximum value and then remain constant at this value for all lower temperatures. This theoretical curve based on the London theory is shown in Fig. 7 for a film of 0·80μ thickness. Comparing this to the curve obtained experimentally,

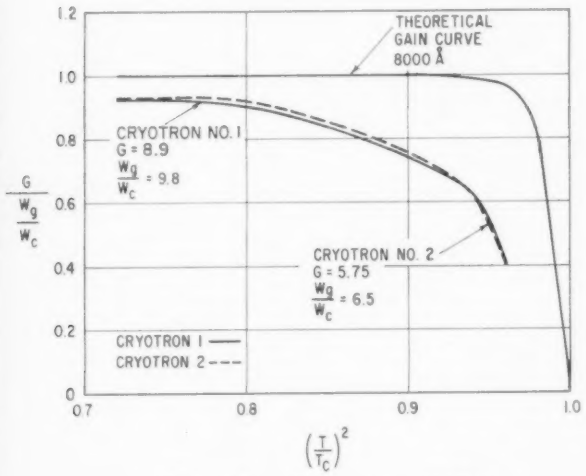


FIG .7. Normalized gain vs. $(T/T_c)^2$.

it can be seen that it reaches a maximum much before the experimental curve. The explanation of this discrepancy is incomplete, but is undoubtedly related to the discrepancy between measured and theoretical field values and also to the discrepancy between the effective thickness and the optically measured thickness of the films[2].

Fig. 8 shows cryotron-characteristic curves taken on one cryotron at two different temperatures. These curves are the result of recording the control current required to restore resistance for many values of gate current between zero and critical. The gate is superconducting for any

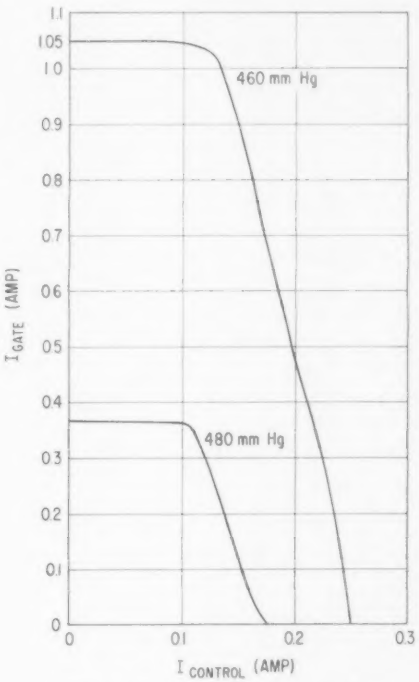


FIG. 8. Typical cryotron gain curves.

combinations that fall between the curve and the axes. The control current intercepts I_{co} of these two gain curves agree with the values predicted from the measured H_c' and the geometry according to

$$I_{co} = H_c' W_c$$

and the gate current intercepts I_{gc} agree closely with the values predicted from H_c and the geometry according to

$$I_{gc} = H_c W_g$$

Since the fields due to these two currents should add in quadrature, the gain curve might be expected to be elliptical in shape with intercepts at I_{gc} and I_{co}. Several explanations have been proposed to account for the deviation from the elliptical shape, especially the flat top. Any physical defect in the gate such as a nick in the gate or a

poor superconducting contact that is not in the region of influence of the control line would cause a flat top but should also cause I_{gc} to be low. However, since the measurement of W_g is taken on the narrowest portion of the gate, both agreement with the predicted I_{gc} and the flat gain curve would be expected for the case where the gate is wider at the crossing than at the position where the measurement was taken. Another possible cause is a double ground-plane effect on the tin gate just at the crossing, which results from positioning the gate between the two hard superconductors—the ground plane and the control. This double ground-plane effect tends to make I_{gc} higher at the crossing than it is over the remainder of the gate length. The portion of the gate under the control line should exhibit some double ground-plane effect while the remainder of the gate is singly shielded. As a result, the gain curve rises more sharply than expected until the singly shielded portion of the gate limits the current-carrying capacity causing the flat top on the gain curve.

MINIATURE-CRYOTRON STUDIES

In order to show the high density potential, some work has been done to show that the characteristics of cryotron circuits established for large cryotrons (125-μ control widths and 750-μ gate widths) also hold for cryotrons with conductor widths reduced by a factor of approximately 10. The control width was 12·5μ and the gate width was 50μ. A normal-size cryotron was included so that a direct comparison could be made between the characteristics of the different-size devices made during the same evaporation process. The characteristics compared quite favorably when normalized to unit-conductor width, and gains of 1·3 were achieved for the miniature cryotrons. Fig. 9 illustrates the miniaturized nature of this device by comparing both the gates and the control to a human hair.

FLIP-FLOP STUDIES

The planar cryotron flip-flop was selected for the first cryogenic switching-circuit studies because of its extreme simplicity from both an operational and analytical viewpoint. The general configuration of the flip-flops tested is shown in Fig. 10. The main gates in branch 1 and branch 2 are made of evaporated tin. They are 500μ wide and from 0·5 to 0·8μ thick. The sense cryotron gates, sense 1 and 2, are each comprised of four tin gates connected in series to increase the sense voltage. Each of these sense gates is 375μ wide and they are the same thickness as the main gates. The main controls, drive 1 and drive 2, and the flip-flop loop control lines (over the sense gates) are made of lead and are 125μ wide. The control lines and ground-plane films are at least 1·0μ thick.

A direct current I_{FF} is applied to the flip-flop loop. This current must be less (but not necessarily much less) than the I_{gc} of the input gates in each branch of the device. If the entire flip-flop loop is superconductive when I_{FF} is applied, the current will divide between the two branches in a manner which is inversely proportional to the inductances of the two paths. When drive 1 is energized with enough current to generate H_c' for gate 1, the gate will undergo a superconducting-to-normal transition, thus causing all of I_{FF} to switch to branch 2. Drive 1 may now be relaxed without disturbing the current flow in branch 2. Any redistribution of current flow would cause a net change in magnetic flux through the superconducting flip-flop loop which would violate a basic law of superconductivity. The current in branch 2 should now be sufficient to restore resistance in the sense cryotrons (sense 2). The sense cryotrons are used to monitor the current in the two flip-flop branches. By adding a bias current (bias connections are shown in Fig. 10) through the hard superconducting portion of the flip-flop loop, it is possible to calibrate the sense-cryotron output voltage very accurately in terms of the flip-flop current in a given branch. The flip-flop current was all switched to branch 2 and monitored on sense cryotron 2. In order to switch the current back to branch 1, drive 2 must be energized. When resistance is inserted in one branch, the current is expected to transfer to the other branch according to the relationship $i = I_{FF}[1-\exp(Rt/L)]$; where R is the resistance (approximately 2·5 mΩ) inserted in the gate, and L is the total inductance of the loop. The resistance switched into the input gates, $Rs_1 = Rs_2$, is measured by applying a small sense current (1 mA) to the flip-flop and then driving the two drives in series. The resulting resistance, E_{FF}/I_{FF}, is the resistance switched into the two branches in parallel so the

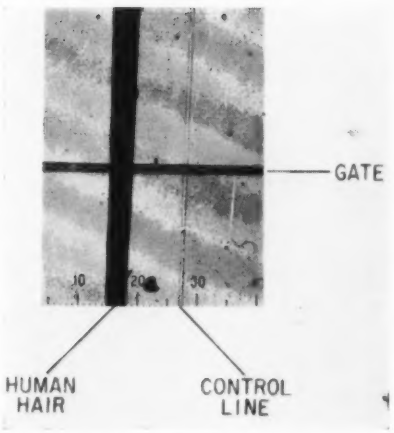


FIG. 9. Miniature cryotron.

[*facing p.* 394

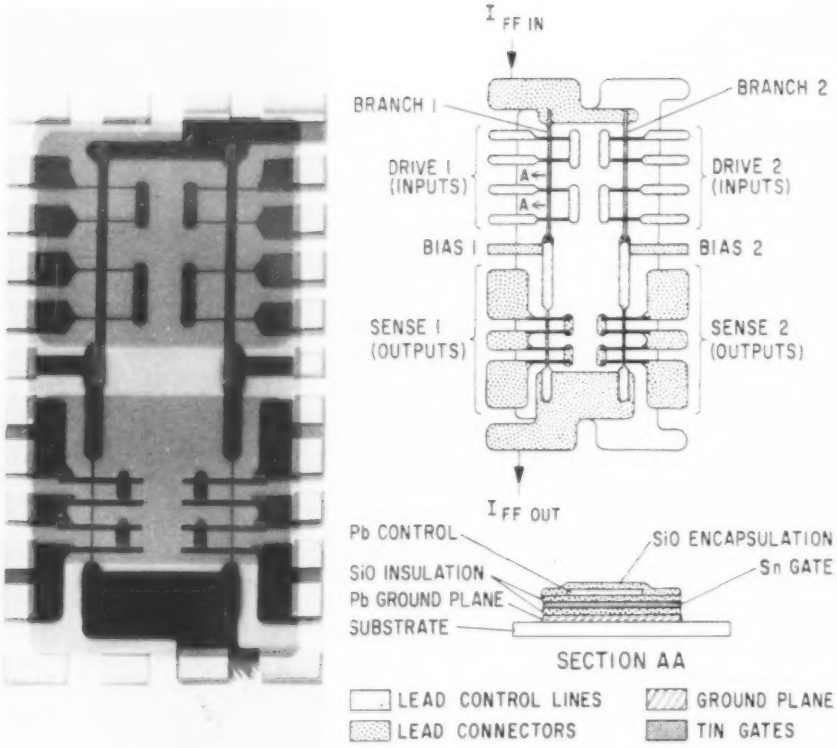


FIG. 10. Experimental flip-flop.

resistance per side is considered to be twice this value. The inductance of a superconducting film over a ground plane is calculated by using the strip transmission-line inductance formula.

FLIP-FLOP TEST RESULTS

The d.c. characteristics of the individual cryotrons comprising the flip-flops were measured using the techniques similar to those described previously. It suffices to say that the flip-flop cryotron characteristics compared favorably with those discussed previously and that H_c' and I_{co} fell well within the tolerances listed.

After obtaining the d.c. characteristics, often one of the first tests performed on the flip-flops is the storing of a persistent current. Persistent currents can be monitored directly if they are equal to or greater than I_{co} of the sense cryotrons. If the persistent current is less than I_{co} of the sense cryotrons, bias current must be used to help drive the sense cryotron.

If the persistent current is allowed to flow for a long period of time and no decrease in the magnitude of the current can be detected, it provides a strong indication that the flip-flop loop has zero resistance, or at least that any resistance in the loop is so small that it is inconsequential. This persistent-current test was performed primarily to insure that all intermetal connections in the flip-flop loops were superconducting. In a typical test a persistent current was allowed to flow for 1 hr, and there was no detectable decay in the persistent-current magnitude for that time period. The results of this test for a flip-flop with an inductance of 1·05 mμH shows that the resistance, if there is any, is less than

$$\frac{L}{T} = \frac{1050 \times 10^{-12}\ \text{H}}{3600\ \text{sec}} = 0{\cdot}3 \times 10^{-12}\ \Omega$$

Since the L/R time constant of the flip-flops is of major interest for operational-speed considerations, the measurement of L/R is discussed in some detail. The most straightforward method of measuring L/R utilizes the sense-cryotron switching characteristics shown in Fig. 11. A value of I_{FF} is chosen so that when all of I_{FF} is forced to the side being sensed, the sense voltage V will vary sharply with a small change in I_{FF}. This point can best be demonstrated by examining the characteristics shown in Fig. 11. I_{FF} equal to I_M (e.g. 63 mA) gives an output voltage of 30 μV in the example. A variation of $\pm$ 2 mA in I_{FF} causes V sense to vary $\pm$ 10 μV so that a very sensitive

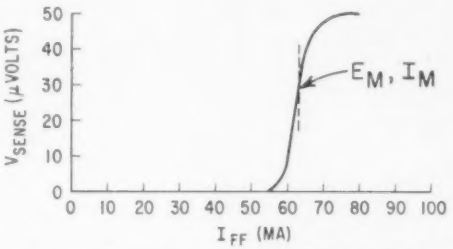


FIG. 11. Sense cryotron switching characteristics.

measurement of I_M (in this case 63 mA) is possible. If I_{FF} is now increased to 100 mA, the 30-μ V sense output becomes a measure of 63 per cent of I_{FF}, this 63 per cent being approximately the amount of current that will switch in one L/R time constant. A push-button pulsing technique is then used to find the pulse width that switches enough current to make V sense exactly 30 μV. The duration of this drive pulse is equal to L/R.

In the actual circuit the bias connection makes it possible to shift the switching characteristic of the sense cryotron (Fig. 11) with respect to the magnitude of flip-flop current, so that the value shown as zero flip-flop current may represent a certain percentage of the I_{co} sense already applied. The calculated inductance, along with the measured value of resistance switched in, can be used to calculate L/R. The variation between the calculated and directly measured values of L/R was less than 10 per cent for ten flip-flops tested.

Another interesting test is the measurement of the current buildup in a flip-flop branch as a function of time. Calibration of a particular value of V sense (e.g. sense 1) as a function of I_{FF} is again required. I_{FF} is then increased and forced into branch 2. The pulse width required to switch the predetermined value of I_{FF} to branch 1 is recorded. If bias current is then introduced, a smaller (or larger) amount of I_{FF} can be switched to branch 1 to give the desired output voltage and consequently a narrower (or wider) pulse will be required. Determining the pulse widths required to produce the same predetermined output

voltage for a wide range of bias currents makes it possible to plot the current switched as a function of time, as shown for a typical flip-flop in Fig. 12. The curve labeled $I_{FF}[1-\exp(-0{\cdot}45t)]$ in Fig. 12 is an exponential curve drawn as a function of the measured L/R time constant of the same flip-flop.

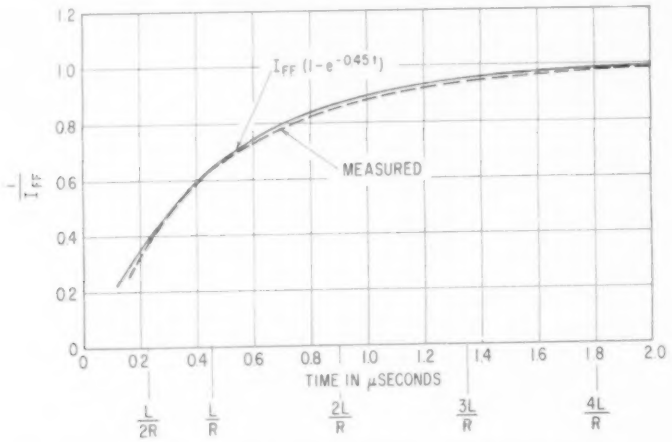


FIG. 12. Per cent current switched as a function of time.

The agreement between the measured curve and the true exponential is within the experimental accuracy. An interesting conclusion from this test is that the resistance switched in is apparently constant during the switching interval so that the resistance does not vary greatly as a function of the current through the gate.

DYNAMIC TESTS

A more revealing flip-flop test is to switch I_{FF} back and forth between the two branches by applying current pulses alternately to the two drives. The input-pulse amplitude should be at least I_{co} so that the resistance switched in will not be affected when the gate current is reduced. In order to get a comparison between repetition rate and the time constant, 63 per cent of I_{FF} should be built up in a branch before the switching direction is reversed. This requires a drive-pulse width equal to one L/R time constant, and it means that the sense cryotron must be driven normal by approximately 60 per cent I_{FF}. The frequency at which resistance was no longer switched into the sense cryotron is listed in the table below. If there is no heat limitation, this frequency should be equal to the reciprocal of the time constant. Of six flip-flops tested, with L/R time constants of more than 1 μsec, the repetition-rate limit was consistently close to the limit set by L/R. This gives good indication that heat buildup is not a problem at frequencies less than 1 Mc/s.

This work has been extended to some fast flip-flops with L/R time constants between 100 and 450 μsec. The decreased time constants were obtained by decreasing the insulation thickness and by increasing the number of flip-flop input crossings. A list is given below of the L/R measurement and the repetition-rate limit obtained on four flip-flops.

Table 1

FF no.	Time constant (mμ sec)	Maximum input-pulse repetition frequency (Mc/s)	
		Calculated	Measured
1	450	2·2	1·6
2	250	4	3*
3	200	5	4
4	100	10	7·5*

* Drive limitations rather than *FF* limit

These results are encouraging. The fact that the repetition-rate limit is consistently lower than the

calculated limit is not too disturbing in view of two testing problems which became more bothersome at increased frequencies. The requirements placed on the rise time of the current drivers used to energize the control lines becomes much more stringent at higher frequencies. When the rise time becomes an appreciable fraction of the pulse width, the test becomes less significant and the measured repetition-rate limit will be well below the calculated value.

Another limitation encountered in testing fast flip-flops was caused by the minute output signal (200 μV). The slower flip-flops permitted use of a high-gain, low-band-pass amplifier, thus giving a signal amplitude readily displayable on an oscilloscope. The sense signal was integrated by these amplifiers so the amplifiers served to filter out the high-frequency noise coupled from the drive lines. Even though the signal-to-noise ratio was very poor, the signal could be recognized quite readily. As the operating frequency of the flip-flops increased, the noise increased, and at a 1 Mc/s repetition frequency the signal-to-noise ratio was less than unity. In order to increase the signal the design of the sense cryotron was changed. Instead of the sense gate being perpendicular to the control line, it was made parallel to the control line, so that resistance was introduced in a much longer length of gate. This increased the sense-signal output by a factor of 10 for the particular geometry used, and permitted observation of circuit repetition frequencies up to 7·5 Mc/s. The flip-flops listed above had this type of sense circuitry.

Fig. 13 shows operation of flip-flop 4 at a 4-Mc/s input frequency. Fig. 13(a) shows the current waveforms measured at the point where the drive coaxial cable enters the dewar. Tests showed that there was little difference in the waveform applied to the flip-flop and the waveform monitored at this point. Fig. 13(b) shows the noise on the sense line (no sense current applied), and 13 (c) shows the output signal of the flip-flop plus the noise.

CONCLUSIONS

Speeds up to 7·5 Mc/s and circuit time constants as low as 100 mμsec have been discussed in some detail. By extending the technique used here, circuits with measured time constants of less than 10 mμsec have been built. A flip-flop of this type was observed switching at a 20 Mc/s input repetition rate. The measured switching time indicates that the device is capable of a 100 Mc/s repetition frequency, but energy from the drive lines couples into the sensing circuit and has prevented observation of switching at this frequency. These speeds have been achieved using thick films (0·70 μ). Thick films were used because good predictability and reproducibility have been demonstrated for cryotrons made with films of this type. As the evaporated-film technology advances so that thinner films with desirable characteristics can be made, circuits with time constants less than 1 mμsec should become feasible.

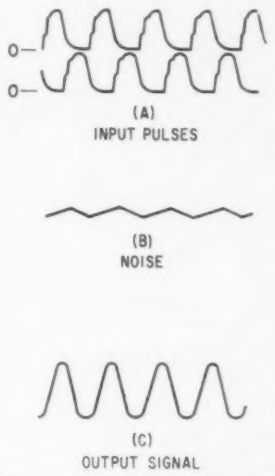


FIG. 13. 4 Mc/s input frequency.

The cryotrons used for the basic-cryotron and flip-flop studies were relatively large (125-μ controls, 750-μ gates); however it has also been demonstrated experimentally that the relationships established through these studies are valid for cryotrons with substrate area reduced by a factor of 100 (12·5-μ controls, 50-μ gates). Cryotron densities of 40×10^6/ft^3 should be possible using cryotrons with these dimensions.

Acknowledgements—Much of the magnetic-field switching work reported was performed by A. STALEY. All mem-

bers of the Low Temperature Development Group under W. J. FITZGERALD contributed to these efforts, especially J. P. BEESLEY, with advice on device development, D. S. WEED with advice on device fabrication, and Dr. W. B. ITTNER, III of IBM Research with suggestions concerning this paper. Much of the fabrication and testing was done by J. D. TORMEDIS, J. J. VAN STEENBURGH and R. W. BURKHART. The author wishes to extend thanks to all of these people.

REFERENCES

1. M. E. BEHRNDT, R. H. BLUMBERG and G. R. GIEDD, *IBM J. Res. Dev.* **4,** 184 (1960).
2. S. TOLANSKY, *Multiple Beam Infterferometry*. Clarendon Press, Oxford (1948).

Solid-State Electronics Pergamon Press 1960. Vol. 1, pp. 399–408. Printed in Great Britain

CRYOTRON STORAGE, ARITHMETIC AND LOGICAL CIRCUITS

M. K. HAYNES
International Business Machines Corporation, Poughkeepsie, New York

INTRODUCTION

A MAJOR advantage of cryotron elements is that storage of digital information may be accomplished by the use of persistent supercurrents. Since the resistance of an element in the super state is identically zero, one may then store a circulating loop of current without any loss due to dissipation of energy by resistance. This provides a type of storage which is unique to computing machinery.

Previous types of storage devices (other than wheels and latches) have fallen into one or the other of two classes: either they have been of the type which give a d.c. indication of their state, or they have been of the type which require no steady-state dissipation of power to maintain the stored information. Examples of the first type are vacuum-tube or transistor flip-flops, which give a d.c. indication of state and which do require d.c. power for maintenance of information. An example of the second type is magnetic-core storage, which requires no steady-state power dissipation for maintenance of the information but is incapable of giving a steady-state indication of the stored state. The cryotron persistent-supercurrent element is the first electronic device capable of combining these two features, both of which are obviously desirable. It is possible to employ persistent-supercurrent techniques in logical and register circuits, in combination with cryotrons as logical elements. Thus one achieves intermediate storage at no additional cost, since the loops in which the information is stored are usually required to be present anyway.

Another major advantage of the cryotron element is its combined simplicity and versatility as a logical element, and the ease with which complicated logical functions can be synthesized with its use. The elementary cryotron functions as an element whose output is of the form "A and not B" where A represents the possibility of current flow through the gate element in the absence of control current, and B represents the presence of current in the control element. This logical function is equivalent to that performed by a normally closed relay contact, where A represents the possibility of current flow through the contact if the relay is unoperated, and B represents the presence of current in the relay coil. This function is clearly a universal element in that any logical function whatever can be readily constructed by its use.

CRYOTRON STORAGE CIRCUITS

While storage may be accomplished by the use of two cryotrons in a cross-coupled flip-flop, this is unnecessary and undesirable. It appears that persistent-supercurrent storage is superior. A two-dimensional storage selection system is probably best, since it minimizes the length of the selection lines running through the storage array, and a cryotron decoding selection system can readily be achieved. The minimal circuit which we have found is shown in Fig. 1. This employs three cryotrons per bit of storage, and is capable of simultaneous and separate access for reading and writing. The three cryotrons A, B and C and their associated connections comprise one memory cell in which one bit of information can be stored.

A "1" can be stored in the memory cell by a persistent current in the loop *abcd*. To set up this persistent current, current is applied to line D and the write select line E is pulsed. The current on line E causes cryotron B to become resistive and the input current on line D is forced to travel the path *abcd* and then out. If current is removed from line E the current in line D will remain in the same path traced. If the current is terminated on line D, then a circulating persisting current will flow in the loop *abcd*. To sense the presence

of a "1" or a "0" in the memory cell, current is applied to the line *F*, the sense line, and to the read select line *G*. Current in line *G* will cause cryotron *A* to become resistive. Thus any current through the memory cell must flow through the path *hgfe* and then out. If there is a persistent current in the loop comprising *abcd* the cryotron *C* will be resistive and, if cryotron *A* is also resistive, no current can flow out on the sense line *F*. The sense current will then be diverted out of the sense line into some other superconducting path connected in parallel with the sense line. Thus, if current appears on line *F* as a result of a read operation on word no. 1, this current must flow through cryotron *C* and indicates that there is a "0" stored in the memory cell. If no current appears on line *F*, there is a "1" stored in the memory cell. Since the access means can be completely separate and independent, simultaneous reading and writing can be done in this memory. For example, it is possible to read word number 1 at the same time that word number 2 is written.

By appropriate connection of the sense line and the input line, as in Fig. 2, it is possible to transfer information from one location to another within storage. Assume that it is desired to transfer the information in word position number 1 to word position number 2. Also assume that there is a "1" stored in a persistent-current loop *A* and a "0" stored in the persistent-current loop *B*. Current applied to the reset line forces the current from source *C* to flow in line *D* and the current from source *E* to flow in line *F*. The current on the reset line is then removed and current is applied to the read select line *G*. Cryotron *H* is driven resistive, blocking the current in line *D*. Cryotron *J* is resistive because of the current flowing in storage loop *A*, and the current in line *D* will be shifted into line *K*.

Cryotron *L* is resistive, forcing the current in

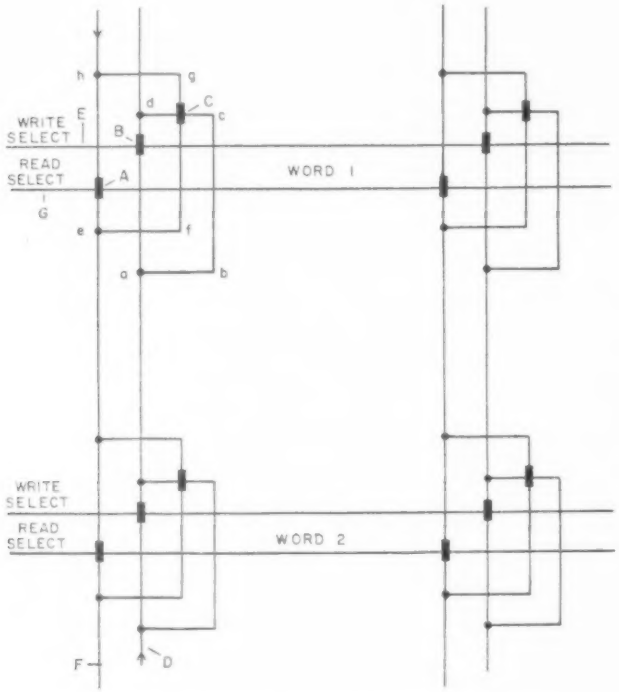


FIG. 1. Two-dimensional persistent-supercurrent storage.

line F to flow through the gate of cryotron M. This is possible because there is no current in the loop B. Current is now removed from line G and applied to the write select line N. The current in line K will now be shunted through line P and, when the current is removed from line N, will continue to flow through line P. When the reset current is applied, the current in line P will form a persistent circulating current. In this manner, a "1" in storage loop A is transferred to storage loop P. The "1" is left in storage loop A. The write select line Q can be pulsed, after a reset pulse, to erase information in word position number 1. Because there was a "0" in loop B, no current is shunted into line R and no current is set up in loop S. If loop S had current in it from prior data, the current would be erased when the write select line N is pulsed.

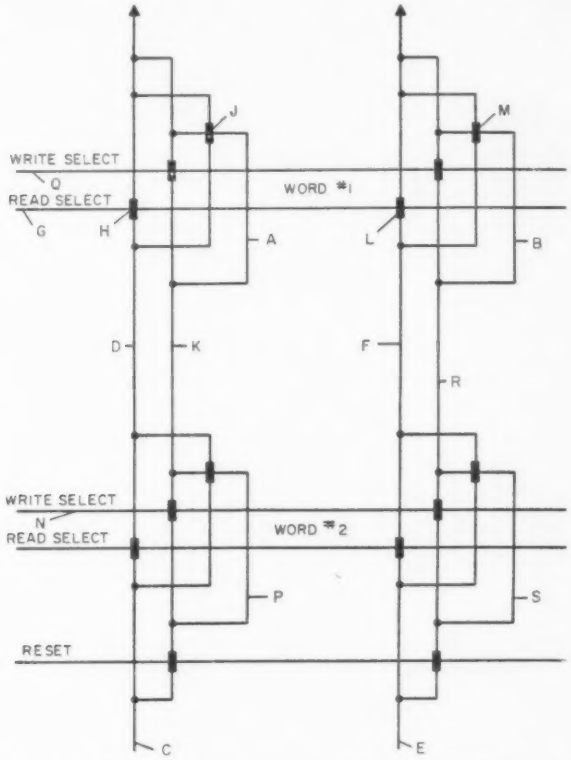


FIG. 2. Persistent-current storage with register-to-register transfer.

By adding two more cryotrons per bit of storage, so that the total is five cryotrons per bit, we are able to achieve simultaneous triple access, as shown in Fig. 3. With this system one can simultaneously read the contents of any two words in storage, and write information into any word in storage, without any restrictions or interlocking between the independent access mechanisms. Such a storage could be used directly to achieve three address operations in connection with the logical unit. That is, any two operands could be fetched simultaneously, and the result of the logical operation upon them can be stored back in memory independently from the operation of the fetch access mechanisms. Other possible uses would be in a machine system having multiple logic units, which could simultaneously read information from a given storage unit without the necessity

for queuing and interlocking problems arising. Another application would be in connection with output equipment which could get simultaneous access to a memory without interfering with or even interrupting the operation of high-speed arithmetic or logical units.

current loop shown in Fig. 4, previously described in reference to Fig. 1, with the eight terminals labeled as shown. An example of more complicated manipulation in storage is shown in Fig. 6, which shows the necessary connections for performing right shifting in storage. By applying

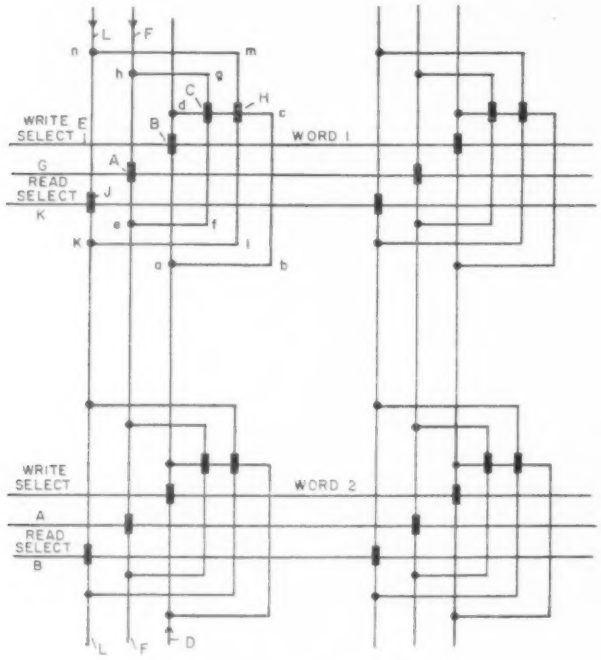


FIG. 3. Triple-access storage.

The memory cell is the same as Fig. 1 except for the additional sense loop *klmn*. An additional cryotron *H* permits control of this additional sense loop by a persistent current stored in the loop *abcd*. A second additional cryotron *J* enables the selection of the additional sense loop by a second read select line *K*. The cryotron *J* controls current in line *L* to sense the presence or absence of a "1" stored in the persistent-current loop. By wiring the second read-access means in a perpendicular fashion, one can readily obtain serial reading of the data, and thus accomplish parallel to serial conversion.

For convenience, let us use a block diagram representation as in Fig. 5 for the basic persistent-

the reset pulse, and then read and write pulses to a given word in memory selected by an address decoder, it is possible to shift the contents of any word in storage one stage to the right per cycle of such applied pulses. By adding circuitry to vary the interconnections, it is possible to control whether shifting is to the right or to the left, and to control the number of places shifted per cycle. Alternatively it would be possible, by applying the write/read pulses to different locations in storage, to cause the information contained in one register to be shifted and then stored in some other register, meanwhile retaining the contents of the originating register location undisturbed. Thus it appears possible to perform right or left shifting

in any reasonable number of stages within a given word in storage or into another selected location in the storage unit.

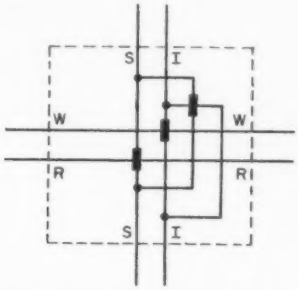


FIG. 4. Persistent-supercurrent storage loop (schematic).

A further application to more complex functioning is shown in Fig. 7, illustrating a method of counting in storage. This shows three successive binary places, lowest order to the right, of two locations of a storage unit which is capable

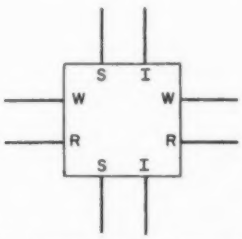


FIG. 5. Block diagram for persistent storage loop.

of adding one to the contents of any memory location and of restoring the incremented count either to the same or to another chosen memory location. The cycle of operation is similar to that of the previously described circuits. The reset pulse is first applied in order to restore the condition of the sense loops to the zero state.

The read pulse is then applied to the word in memory whose contents are to be incremented. This causes the contents of this memory location to appear in the sense lines which pass through the incrementing network shown at the top of the figure. The count pulse, (one, if it is desired to increment the contents of the memory location selected, zero, if it is desired to rewrite the contents unchanged) can be applied concurrently with either the read pulse, the write pulse, or both. It must not, however, overlap the reset pulse.

The incrementing network shown performs the function of complementing the state of the storage

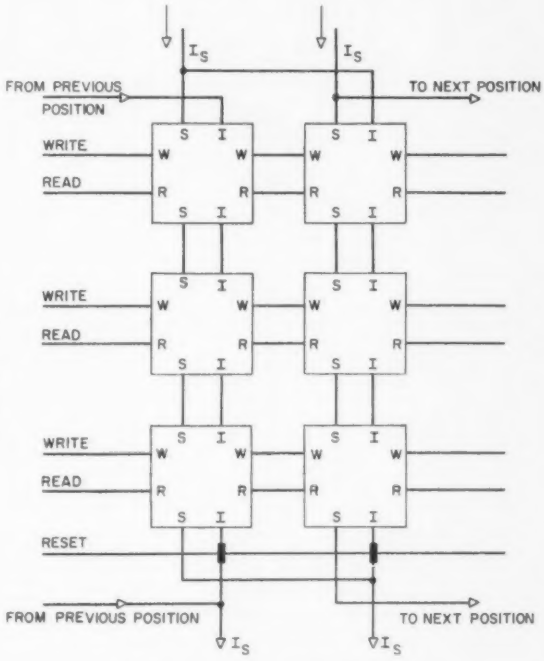


FIG. 6. Right shifting in storage.

cell when a count pulse is received from a lower-order storage position, and transmits a count pulse to a higher-order position whenever a count pulse is received and the previous state of the cell is a one. After the read pulse, the write pulse is applied which causes the incremented value to be stored in the location designated. If the read and write locations are not the same, the contents of the location read are retained undisturbed.

CRYOTRON ARITHMETIC AND LOGIC CIRCUITS

It is possible to use the same persistent-current loops in data registers, in interchange of information between registers, and shifting of registers. These various circuits use the unique features of cryotrons. They achieve the storage function entirely by persistent supercurrents. The temporary storage required in the logic, transfer and shifting

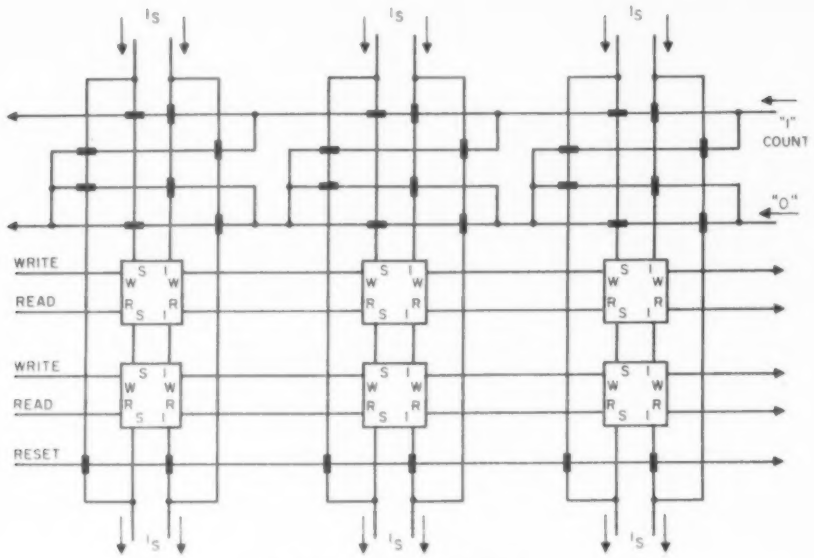


FIG. 7. Counting in storage.

is accomplished at no cost in cryotron count, since temporary storage can be achieved by persistent supercurrents in the loops which interconnect the various storage cells. A typical nonmemory application for these storage loops, a shifting register, is illustrated in Fig. 8. This shift register requires only one storage cell, or block, per bit of information stored. Its operation requires the application of three consecutive pulses to shift the information in the register one place to the right.

The first pulse is applied to the reset line. This has the function of resetting the state of the interconnecting loops of sense and input lines to the zero state, so that wherever a zero state is read in a given cell, a zero state will then be written into the next storage cell to the right.

Next follows the read pulse, whose function is to sense the state of the persistent supercurrent storage loops in each of these cells, or blocks. Wherever the stored state is a zero, or no circulating current, this causes no change in the distribution of currents in the interconnecting loops of sense and input lines. Wherever a one state is sensed in the persistent storage cell, this causes the introduction of a resistance in series with the

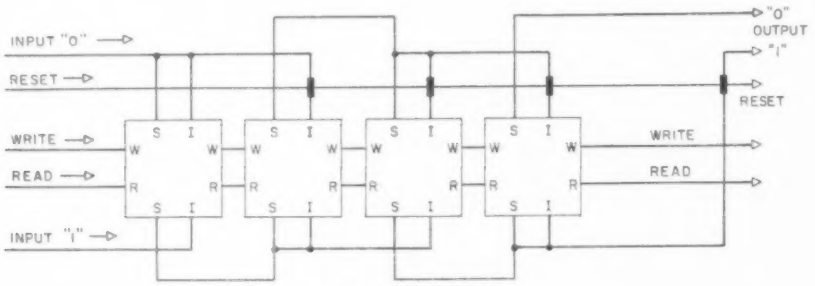


FIG. 8. Shift register.

corresponding sense line, which in turn causes the current it carries to be diverted into the input line of the next storage cell. When this condition has been stabilized, the read pulse may then be removed. The stored state then consists of the original pattern of bits contained in the storage cells themselves, and this pattern is duplicated in the distribution of currents between sense and input lines of adjacent storage-cell positions.

The last pulse to be applied is the write pulse, which then causes the pattern of bits, in the current distribution in interconnecting wires, to be written into the persistent storage cells, shifted one place to the right from its original position. When the new stored states have become stabilized, the write pulse is removed and the cycle may be repeated at will. Note that after the write operation has been completed, the distribution of currents in the interconnecting loops is unchanged, thus necessitating the application of the reset pulse before the next read operation can be performed. By paralleling additional cells with those shown in Fig. 8, this system can be generalized to the shifting of characters or of binary words in parallel through an ordered sequence of storage locations.

In Fig. 9 is shown another application. Here are shown elements of two registers, an upper register and a lower register, whose data contents may be interchanged. As before, the reset pulse must be applied before reading in order to insure a zero-state representation in the interconnecting-loop current distribution. On application of the read pulse, the state of each cell is sensed and the corresponding state is established in the interconnecting loop associated with it. After the read pulse is terminated, the write pulse may be applied, causing each cell of the upper register to store the bit previously contained in the lower register and vice versa. This circuit may also be viewed as similar to that of Fig. 8, but with the loop of the shifting register closed upon itself in a pattern of two's.

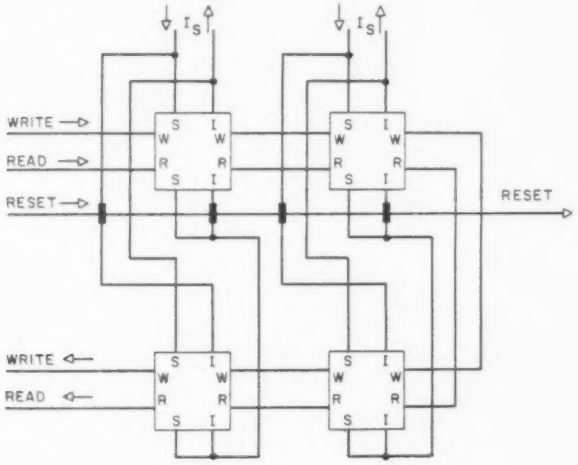


FIG. 9. Data interchange between registers.

In the process of studying logical circuits, we have found that enormous simplifications are possible, from previous ordinary circuit design, by taking advantage of the unique characteristics of cryotrons. By using the facility of denial of input variables as well as denial of the output function, and the ease of permuting input variables, it has been found possible to construct logical circuits from a far smaller group of basic circuits than has previously been possible. For example, any of the sixteen possible functions of two binary-input variables can be achieved by the selection of only one out of two basic circuits, plus the appropriate input and output connections. Similarly for the three binary-input variables case, there are 256 possible functions of three input

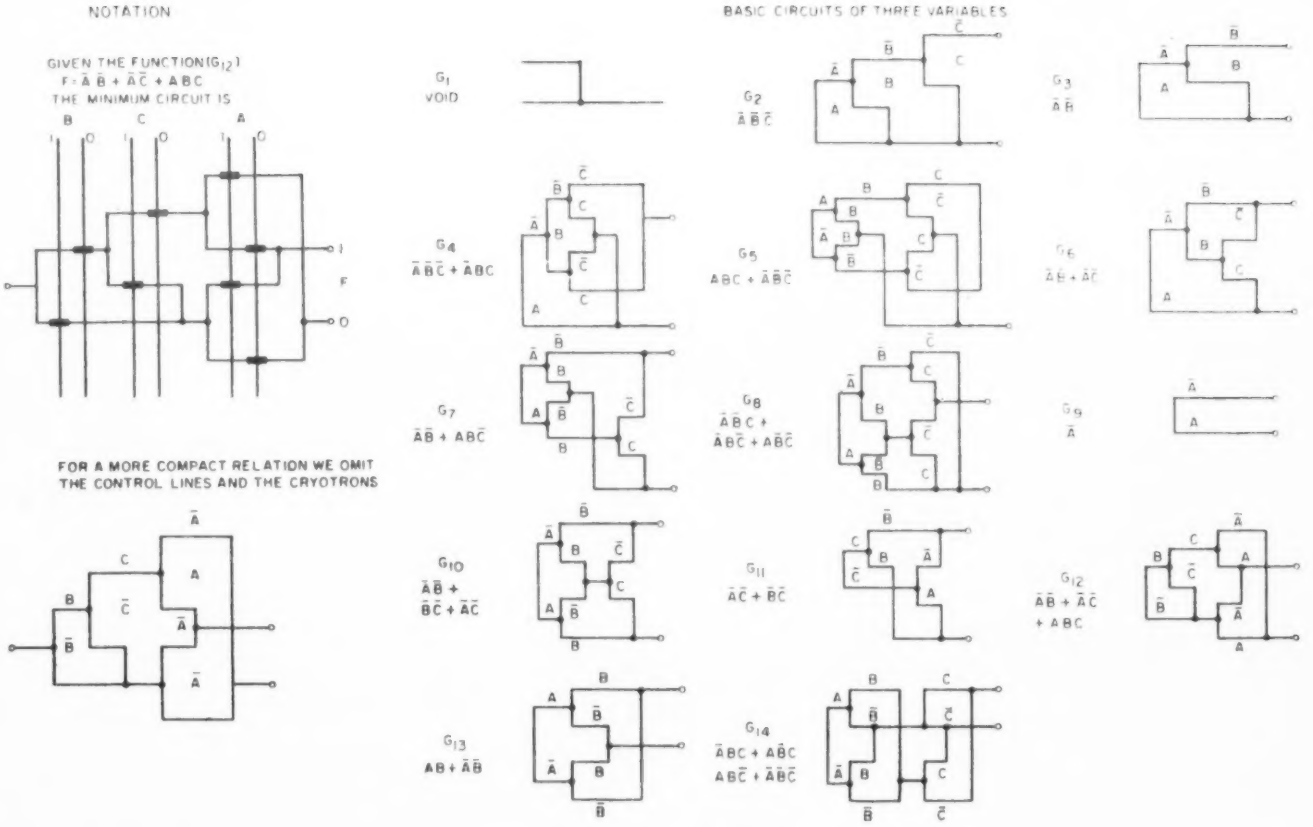


FIG. 10. Basic circuits of three variables.

variables; but these 256 functions reduce, by use of denials and permutations, to only ten basic circuits of three full input variables. Similarly for four input variables, the 65, 536 possible functions reduce to only 221 basic circuits from which a design choice need be made.

Any possible three-variable function can be achieved in a "collapsed-tree" structure with no more than ten cryotrons. All of these three-variable circuits, including those of fewer variables, are shown in Fig. 10. Any possible four-variable function can be similarly realized with no more than sixteen cryotrons.

A similarity of cryotron logic to relay logic is that the elements may be employed to accomplish "transmission logic". One can set up a network of cryotron gate elements by causing the appropriate control currents to flow through the control elements, in a manner analogous to the situation where one picks or drops out desired relays. Then a probing current is applied to the entire network and allowed to propagate as rapidly as possible through the network of cryotron elements, thus yielding an output indication in a minimum time determined by the transmission delays of the network, not hindered by the cascading of element transition times or threshold delays. This transmission-logic operation is important to a number of useful cases such as that of a parallel adder or a high–low–equal comparison.

A parallel binary full adder can be realized with fourteen cryotrons, as shown in Fig. 11. This adder is self-timing, in that the lowest-order carry inputs can be so controlled as to initiate the add operation,

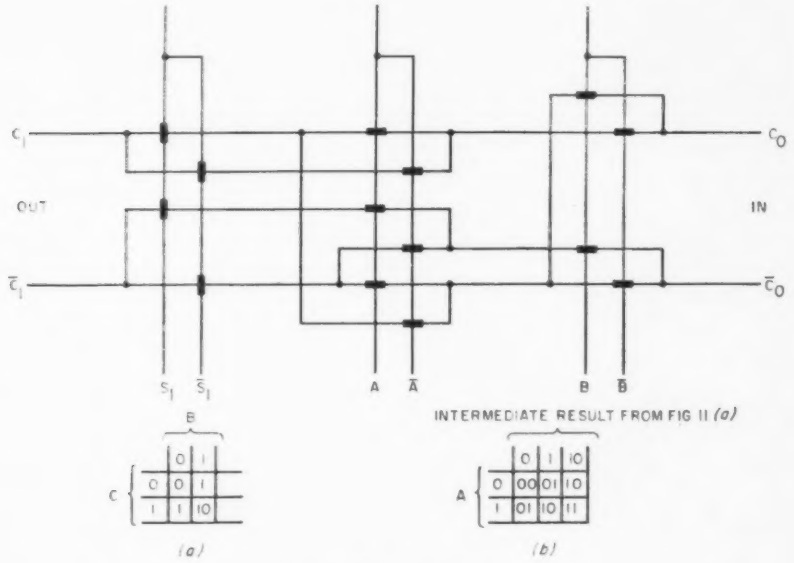


FIG. 11. Binary full adder.

and carry completion can be determined by sensing the highest-order carry outputs. The carry propagation chain does not include cascaded threshold discrimination operations which would limit the carry propagation speed. Furthermore, neither the addend nor augend inputs are used as current sources to the adder, thus imposing no restrictions on the form of availability of the input circuits.

The two lines C_0 and $\bar{C}_0$ are the carry inputs to the stage. If the carry input is a "1", there is current on the C_0 line and no current on the $\bar{C}_0$ line. If the carry input is a "0", there is current on the $\bar{C}_0$ line and no current on the C_0 line. The carry input is first operated on by the B inputs, and, referring to Fig. 11(a), it is seen that there are three possible results from the operation: 0, 1

and 10. The results of the first operation can be recombined into three lines, and these three lines can be considered as carrying an intermediate result which must further be operated on by the A inputs. Referring to Fig. 11(b), the four possible results of this operation are 0 0, 0 1, 10 and 11. A further recombination is thus possible to four lines which generate the sum output on line S_1 or $\bar{S}_1$, and then recombine into two lines C_1 and $\bar{C}_1$ to form the carry output.

SUMMARY

Cryotron elements have a number of characteristics which are unique amongst those of other control, storage and logic devices. These are: absolute-zero resistance of the superconducting state, the concomitant lack of attenuation of propagating signals, persistent-current storage, simplicity and versatility in logic, low power and energy levels, small size and projected low cost. These unique characteristics make possible new and simpler forms of storage, arithmetic and logical circuitry which may revolutionize the design of digital data-processing systems.

Acknowledgements—It is a pleasure to acknowledge the support of the U.S. Navy Bureau of Ships, and the assistance of Mr. E. H. SUSSENGUTH in developing the basic circuits shown in Fig. 10.